DOCUMENT RESUME

ED 442 837 TM 031 257

AUTHOR Chang, Shun-Wen; Ansley, Timothy N.; Lin, Sieh-Hwa
TITLE Performance of Item Exposure Control Methods in Computerized Adaptive Testing: Further Explorations.
PUB DATE 2000-04-00
NOTE 44p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Adaptive Testing; Algorithms; *Computer Assisted Testing; Item Banks; Sample Size; *Test Items
IDENTIFIERS *Item Exposure (Tests)

ABSTRACT

This study examined the effectiveness of the Sympson and Hetter conditional procedure (SHC), a modification of the Sympson and Hetter (1985) algorithm, in controlling the exposure rates of items in a computerized adaptive testing (CAT) environment. The properties of the procedure were compared with those of the Davey and Parshall (1995) and the Stocking and Lewis (1995) (SLC) conditional multinomial procedures within the purview of estimating examinee's abilities. Each of the exposure control methods was incorporated into the item selection procedure and the adaptive testing progressed based on the CAT design established for this study. The advantages and disadvantages of these strategies were considered under four item pool sizes and two desired maximum exposure rates and were evaluated in light of test security, test overlap rates, utilization of the item pool, and conditional standard errors of measurement. Also, the issue of the appropriate conditional sample sizes in deriving the exposure control parameters was considered in the present study. Simulation results show no effect of using the four conditional sample sizes. The SHC produced the most satisfactory results in terms of item security and test overlap rates followed by the SLC method. Results also show that as long as the control for item exposure was not exercised, optimal items could be administered to almost every examinee under any of the four item pools. Findings of this study provide useful insights on how item pool sizes and maximum item exposure rates affect the performance of the exposure control methods. (Contains 3 tables, 13 figures, and 18 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
S-W. Chang
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ED 442 837

Performance of item exposure control methods in computerized adaptive testing: Further explorations

Shun-Wen Chang
National Taiwan Normal University
Timothy N. Ansley
The University of Iowa
Sieh-Hwa Lin
National Taiwan Normal University

Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, April 2000.

TM031257

BEST COPY AVAILABLE

Abstract

This study examined the effectiveness of the Sympson and Hetter conditional procedure, a modification of the Sympson and Hetter algorithm, in controlling the exposure rates of items in the CAT environment. Its properties were compared with those of the Davey and Parshall and the Stocking and Lewis conditional multinomial procedures within the purview of estimating examinees' abilities. Each of the exposure control methods was incorporated into the item selection procedure and the adaptive testing progressed based on the CAT design established for this study. The advantages and disadvantages of these strategies were considered under four item pool sizes and two desired maximum exposure rates and were evaluated in light of test security, test overlap rates, utilization of the item pool, and conditional standard errors of measurement. Also, the issue of the appropriate conditional sample sizes in deriving the exposure control parameters was considered in the present study.

The results of this study should provide valuable insights on the issue of item exposure control and also offer a clearer understanding of the properties of the exposure control methods. By incorporating an appropriate algorithm into the item selection process, optimal items will not be overexposed and test security in the CAT environment should be improved.

Acknowledgement

The authors wish to thank Bor-Yaun Twu for his assistance in programming the algorithms.

With the growing number of large-scale applications in computerized adaptive testing (CAT), measurement professionals are encountering many practical issues while the theory is translated into practice. Test security is one of the primary issues that substantially concern the makers of high stakes tests for the continuous testing context of computerized adaptive tests (CATs). The main security issue is that some items of the CATs might be so popular that they appear frequently to test-takers within a short period of time. When test-takers have access to the questions before test administrations, frequently appearing test items may soon be compromised. High rates of item exposure thus lead to a great test security risk.

The issue of controlling exposure rates of the CAT items has been studied through the management of item pools as well as the incorporation of exposure control methods into the item selection process. This study focused on the item exposure control methods that have been proposed to date. The 5-4-3-2-1 method of McBride and Martin (1983) was an early attempt to prevent the overexposure of items, in which items are randomly selected from among a group of items with approximately equal amounts of information at the early stages of the testing process. By arranging the random selection from among a sequence of item group sizes of five, four, three, two and one, the most informative item at a current ability estimate in the early testing process is thus not always administered. However, this procedure has been shown not to ensure item security to any noticeable degree better than a procedure with no exposure control (Chang, 1998).

By embedding statistical mechanisms in the item selection procedure, later methods seek to control the exposure rates of items to a desired maximum value, r, that is specified in advance of testing. The Sympson and Hetter (Hetter & Sympson, 1997; Sympson & Hetter, 1985) (SH) procedure employs an exposure control parameter for each item in the pool, which is determined after a series of iterative simulations. Given that an item has been selected, whether to administer this item to the examinee depends upon the exposure control parameter of this item. For very popular items, the exposure control parameters could be as low as the pre-specified desired exposure rate, indicating that these items cannot be freely administered when they are selected. For items rarely appearing, the associated exposure control parameters could be as high as 1.0, meaning that once these items are selected, they are almost always presented. Items that have been selected but not administered are excluded from the pool of remaining items for the examinee. The probabilistic model of Sympson and Hetter seeks to achieve the goal that there is no item administered to more than a pre-specified fraction of examinees.

The Davey and Parshall (1995) (DP) methodology provides an exposure parameter for each item that is conditioned on all other items previously administered to the examinee. To utilize this method, an exposure table needs to be prepared through a series of simulations. Diagonal elements of the table indicate the probability limits with which individual items can be administered given selection. These

values will be small if the corresponding items tend to be selected frequently. The off-diagonal elements represent the probability limits with which a pair or set of items can appear together given selection. Similarly, the off-diagonal values will be small if the pairs of items tend to occur together very often. It was concluded in Chang (1998) and Davey and Parshall (1995) that this procedure satisfactorily reduces the extent to which the items overlap across tests administered for examinees with similar ability and for examinees of differing ability.

The Stocking and Lewis unconditional multinomial (SL) procedure (Stocking & Lewis, 1995a, 1998) was derived by remodeling the SH approach. This method develops an exposure control parameter for each item following the SH algorithm but it differs in the way an item is selected. Rather than use optimal item selection, Stocking and Lewis employed a multinomial model to select the next item for administration, in which a cumulative multinomial distribution is established by successive addition of the operant probabilities. The operant probability represents the joint probability that all items before a particular item are rejected given selection and that item is administered given it is selected. To guarantee a full length adaptive test with the multinomial selection model, the length of the list of items from which the next item is selected may be formed by dividing the item pool size by the adaptive test length (M. L. Stocking, personal communication, March 24, 2000). The integer of this ratio is the length of the list from which the next item is selected. All items appearing in this list preceding the item actually administered are removed from further consideration in this adaptive test. Stocking and Lewis (1995a) claimed that the SL method retains the advantage of the SH strategy in that it serves to directly control the probability of administering an item when the adaptive tests are presented to the target population of examinees but causes fewer convergence problems than the SH procedure during the iterations of the adjustment simulations.

Stocking and Lewis (1995b, 1998) also proposed the conditional multinomial method to directly control the item exposure to examinees at the same or similar levels of proficiency. Unlike the unconditional methods where an exposure control parameter is developed for an item in order to limit the item's overall appearance in reference to the examinee group, the Stocking and Lewis conditional multinomial (SLC) procedure derives for each item an exposure control parameter with respect to a particular level of examinee ability. Accordingly, this procedure results in different exposure control parameters for each item to be applied to various ability levels.

Investigations on the properties of these strategies showed that the DP and SLC procedures as methods of exposure control were the most effective (Chang, 1998). However, developing the exposure control parameters for the DP procedure and the SLC method was very tedious and time-consuming. Based on the results of the SH procedure that were very similar to those of the SL procedure but with an easier derivation process, Chang suggested that by deriving the SH exposure control parameters in

reference to each ability level, the *Sympson and Hetter conditional procedure* (SHC) might be a competitive method to the SLC method in controlling the item exposure rate. This study attempted to explore the behaviors of the SHC procedure and to compare its properties with those of the DP and SLC procedures within the scope of estimating examinees' abilities in a CAT context.

Based on the results of Chang (1998) that the effects of the item pool sizes and the desired maximum exposure rates on the performance of the procedures were not consistent, the current study is designed to further examine how the exposure control methods are affected differentially by the employment of varying sizes of item pool and varying maximum rates of item exposure.

The issue of the appropriate conditional sample sizes in developing the exposure control parameters for the SHC procedure and the SLC strategy was considered in the current study. As indicated in Stocking and Lewis (1995b, 1998), larger conditional sample sizes would lead to smoother convergence of the procedure to values closer to the target maximum exposure rates. A conditional sample size of about 10,000 examinees at each ability point for developing the conditional exposure control parameters was recommended by M. L. Stocking (personal communication, February 17, 1998) for producing better estimates of the maximum observed exposure rates. There is no doubt that the larger the sample size, the better the estimates of the maximum exposure rates. However, concerning the amount of time that would be consumed in the derivation process, employing such a huge conditional sample size may not be very practical in many settings. It is important that appropriate conditional sample sizes for producing stable estimates of the maximum exposure rates be determined to prevent a tedious iteration process.

The Purpose

Previous research has compared some of the item exposure control strategies (Chang, 1998; Davey & Parshall, 1995; Featherman, Subhiyah, & Hadadi, 1996; Parshall, Davey, & Nering, 1998; Revuelta & Ponsoda, 1998; Stocking & Lewis, 1995b, 1998; see Chang, 1998, for a summary of the literature). This study further explored and compared the advantages and disadvantages of item exposure control algorithms within the scope of estimating examinees' abilities in a CAT context. Specifically, the current study attempted to achieve the following objectives:

1. to examine the effectiveness of a modification of the SH procedure and to offer more information about the properties of the various exposure control strategies;
2. to provide information about how the exposure control methods are affected by item pool sizes and desired maximum exposure rates;
3. to suggest appropriate conditional sample sizes for developing stable exposure control parameters.

Method and Data

Simulations were employed to carry out this research. The plan of the study design and methods for data analyses are described below.

Design of the Study

Specification of Decisions for the CAT Components

The 3-PL model of IRT formed the basis for the investigations of this study. Real pools of discrete items were used, in which the item parameters were calibrated from multiple forms of the existing paper-and-pencil ACT Assessment Mathematics Test (ACT, 1997) using BILOG (Mislevy & Bock, 1990) on a single ability scale. The CAT administration process was initiated by assigning each examinee a common ability estimate of zero to simulate a situation where no a priori information was available about the individuals.

Items were selected based on the maximum item information criterion with the various exposure control algorithms incorporated into the selection process. The content presented to the examinees was balanced according to the Kingsbury and Zara's mechanism (1989). The first item was administered following the algorithms of item exposure control, regardless of the item's content attribute. The percentage of items that had been administered in each content category was calculated and compared to the corresponding pre-specified percentage. The content area with the largest discrepancy between the empirical and the desired percentage was then identified, from which the next item was selected and administered based on the algorithms of the various exposure control methods. To estimate the examinees' abilities, Owen's Bayesian strategy (Owen, 1975) was utilized for the provisional ability estimation and the maximum likelihood estimation method was employed for the final ability estimation. Each examinee was administered 30 items.

Specification of Factors Included in the Study

1. Item Exposure Control Methods

The DP methodology, the SHC procedure and the SLC procedure were the exposure control methods to be investigated. The NO procedure, the item selection with no control exercised over item exposure rates, was included as part of the study for comparison purposes.

2. Item Pool Sizes

Four item pools of 360, 480, 600 and 720 items of very similar quality were used, with the larger item pools subsuming the smaller ones. The mean and SD of the *a* parameters were around 1.02 and .33; the mean and SD of the *b*s ranged from .15 to.19 and 1.04 to 1.08, respectively; and the mean and SD of the *c*s were .18 and .08. In these item pools, there existed more items at the middle of the ability

continuum than at the two ends. Also, all pools contained items that discriminate moderately well with more discriminating power at the middle difficulty levels than at the extreme levels. The effects of enlarging the pool size on the performance of the various algorithms were examined.

3. Desired Maximum Exposure Rates

The two target maximum desired exposure rates of .10 and .20 were studied. That is, the exposure control algorithms maintained the maximum observed appearance rates for the most used items in the pool below .10 or .20.

4. Conditional Sample Sizes

Sample sizes of 4,000, 6,000, 8,000 and 10,000 were attempted for exploring the appropriate conditional sample sizes at each ability level necessary for producing stable exposure control parameters.

Procedures for Data Simulation

For each combination of the exposure control methods, the pool sizes and the desired maximum exposure rates, a sequence of simulations was performed. Within each sequence of simulations, two stages were involved. First, the exposure control parameter for each item in the pool was developed according to the specific exposure control algorithm and the CAT design proposed for this study. Then, simulations were carried out for the operational CAT administrations.

The Stage of Developing Exposure Control Parameters

To determine the exposure control parameters for the items, this stage proceeded as follows.

For the DP methodology, the adaptive tests were administered to a sample of 50,000 examinees drawn from a standard normal ability distribution (i.e., $N(0,1)$) on the theta metric. For the SHC and SLC procedures, the development of exposure control parameters was in reference to a particular level of proficiency. The adaptive tests were respectively administered to the various conditional samples of 4,000, 6,000, 8,000, and 10,000 examinees at each of the theta levels equally spaced over the interval of -3.2 and 3.2 with an increment of .40 (i.e., -3.2, -2.8,..., 3.2), totalling 17 ability points.

During the process of administering adaptive tests, tentative item exposure parameters for the various exposure control techniques were developed according to the specific algorithms. However, one problem was detected in the preliminary iteration tryouts while applying the SHC method to the 360- and 480-item pools under the target exposure rate of .10. The iterations were terminated before the specified iterative number had been reached due to the lack of items for selection. This problem might be caused by the exclusion of the selected but not administered items from the pool of remaining items for an examinee's adaptive test. While computing the tentative exposure control parameters for the SHC procedure, the values of the 30 largest new exposure control parameters were set to 1.0 according to the SH algorithm to ensure a complete 30-item test, but the process was carried out with respect to each

ability level. When some of these items have already been removed from further consideration at ability levels in an earlier testing process, there would exit a fewer than sufficient number of items for selection to guarantee the full test administration. This problem appeared especially when the pool size was small and the desired exposure rate was low in that a sufficiently large number of items would be demanded for developing the conditional exposure control parameters. To overcome this early termination problem in the derivation of the SHC exposure control parameters, the selected but not administered items were reconsidered in this study for the 360- and 480-item pools under the desired rate of .10.

For the SLC procedure, the desirability of items was ordered only based on item information values, not on the weights as specified in the original procedures for which the Stocking/Swanson weighted deviations model (WDM) (Stocking & Swanson, 1993; Swanson & Stocking, 1993) was employed to select items. The length of the ordered list from which the next item is selected was determined by dividing the item pool sizes of 360, 480, 600 and 720 by the CAT length of 30 items. All items appearing in this ordered list preceding the item actually administered were removed from further consideration in the examinee's adaptive test. The process repeated until the observed maximum exposure rates were approximately equal to the desired level and the exposure control parameters were stabilized in the subsequent iterations. The stabilized parameters at the final round of iterations were the exposure control parameters to be used in operational adaptive testing.

The Stage of Simulating Operational CAT Administrations

The adaptive test was delivered to each examinee of a sample of 50,000 simulees drawn from a standard normal distribution, following the CAT design established for this study. Items were selected and administered according to the specific algorithm of an exposure control strategy. The exposure control parameters developed from the previous stage were utilized here to manage the administration frequencies of the selected items. The exposure control parameters of the SHC and SLC procedures derived based on the conditional sample size of 10,000 were utilized for optimal comparisons. The item selection with no exposure control was also applied to the operational CAT administrations.

Also, the adaptive tests were administered to a conditional sample of 3,000 examinees at each of the ability levels equally spaced over the range of interest between -3.2 and 3.2, at intervals of size .40. This step was performed to obtain the conditional maximum observed exposure rates and the test-retest overlap rates, as well as to evaluate the measurement properties conditional on each ability point.

Methods for Data Analyses

The criteria for evaluating the strengths and weaknesses of the exposure control procedures under the various conditions were the item security, the test overlap rates, the item pool utilization and the conditional standard errors of measurement (CSEM) of the ability estimates. The degree of item security

achieved by the various methods was indicated by the maximum exposure rates observed in reference to the entire examinee group and also to the examinees of a particular ability level. The values of the test overlap rates were classified into the test-retest overlap rates and the peer-to-peer overlap rates to show the extent to which pairs of items appeared together across tests taken by examinees of similar ability and differing abilities, respectively. The test-retest mean overlap rate was obtained by first computing the percent of items that overlapped between the adaptive tests given to an examinee of ability theta twice, then averaging the overlap percentages over all examinees at this theta level. The peer-to-peer mean overlap rate was obtained by first calculating the overlap percentage of tests taken by two examinees generated from a $N(0,1)$ distribution, then averaging the overlap percentages over all paired examinees. The utilization of the item pool was assessed using the numbers and/or percentages of items in the pool never administered. The CSEMs were the errors of the ability estimation as a result of introducing the various exposure control algorithms into the item selection process.

Results and Discussion

The Results of Developing Exposure Control Parameters

The exposure control parameters were developed through a series of adjustment simulations for the DP methodology, the SHC approach and the SLC procedure. For each of the procedures under the four item pool sizes and the two desired exposure rates, the maximum exposure rates observed are displayed in figures for each iteration stage to illustrate the converging status of the procedures. For the SHC and SLC procedures, the results of employing the various conditional sample sizes to develop exposure control parameters are presented together in same figures for ease of comparisons. Only on the top left plot of each figure is the legend placed to describe the contents of the graph. Every other graph adopts the same symbol annotation. A horizontal line is contained in each figure to indicate the desired exposure rate of $r = .10$ or $.20$.

The numbers of iterations carried out for the DP procedure and both SHC and SLC methods in developing the exposure parameters were not the same. The decision regarding the iteration numbers for the respective methods that were deemed necessary and appropriate for being able to detect whether the exposure control parameters converged was based on the suggestions made in the original studies of the respective algorithms and some preliminary tryouts employing the CAT procedures designed for the current study. Described below are the results of the iterations to develop the exposure control parameters.

The DP Procedure

The development of the exposure control parameters for the DP method was performed through 80 iterations. As shown in Figure 1, the observed maximum exposure rates converged to a value very

close to the pre-specified rate of .10 or .20. For the desired rate of .10, the iterations converged faster as larger item pools were employed while for the desired rate of .20, the observed maximum exposure rates converged at almost the same stage for the various item pools. It can be seen that when the desired exposure rate of .20 was specified, the exposure control parameters were stabilized with fewer iterative steps than when the rate was .10.

The SHC and SLC Procedures

For the SHC and SLC procedures, the exposure control parameters were derived in reference to a particular proficiency level. The iteration simulations for both procedures were conducted through 30 iterations. Although fewer iterative steps were repeated for these two conditional procedures than for the DP methodology, the amount of time needed for one iteration was greater with the conditional methods since the iteration was performed for 17 ability points. The entire derivation process was very time-consuming, especially for the SLC procedure.

The iteration results employing the four conditional sample sizes for these two methods are presented in Figures 2 through 9, where each figure contains results of employing one particular item pool in combination with a desired exposure rate. As displayed in these figures, the SHC and SLC procedures produced similar patterns of the iteration curves, except for the 360- and 480-item pools under the expected rate of .10. The values of both strategies did not approach the pre-specified values as closely as those of the DP procedure. Also, the conditional observed maximum exposure rates for the various ability levels seem to converge to different values. The closer the iterations were at the extreme ability levels, the higher values the conditional observed maximum exposure rates converged to. These results reflected the nature of the real item pools, in that there were fewer items appropriate for administration for the extreme ability levels than for the middle.

However, unlike the results in Chang (1998) that the unconditional counterparts of the SHC and SLC procedures--the SH and SL methods--produced almost identical iteration curves, differences were observed in the outcomes of these iteration simulations. First, the SLC procedure converged to a higher rate than the SHC procedure, especially when the pool size was small and the desired rate was low. Second, the maximum observed exposure rates converged for the various ability levels seemed more different with the SLC procedure than with the SHC procedure. Third, when the target exposure rate of .10 was specified, the SLC procedure seemed to converge a few iterative steps faster than the SHC procedure for most ability levels.

The reason the SLC procedure failed to yield the maximum observed exposure rates as low as the pre-specified rate of .10 or .20 might be that this procedure selected the next item from a list of items that did not include all available items in the pool. Although results of complete test administration were

satisfactory by selecting the items from the list of appropriate length, the observed exposure rates of items were likely increased since a shorter length of the list from which the items were selected would cause the items to be used more often.

It can be detected that the iteration curves of the SHC procedure in Figures 2 and 3 were somewhat different from the curves of this procedure in Figures 4 through 9. While applying the SHC method to the 360- and 480-item pools under the desired exposure rate of .10, the selected but not administered items were considered again for selection. Using this modified algorithm for these two conditions, the SHC iterations were continued through 30 iterations; however, that appeared to cause some convergence problems during the iterations for some ability levels.

The results of employing the various conditional sample sizes in developing the exposure control parameters for the SHC and SLC procedures can also be seen in Figures 2 through 9. The employment of larger conditional sample sizes did not appear to lead to smoother convergence of the two conditional procedures to values closer to the target maximum exposure rates. It can be detected that employing the conditional sample size as large as of 10,000 at each ability level did not produce better estimates of the maximum observed exposure rates than the size of 4,000. Utilization of a large conditional sample size may thus not be necessary or practical in the development of the exposure control parameters, considering the amount of time that would be consumed in the derivation process.

The Results of Simulating Operational CAT Administrations

The exposure control parameters of the DP, SHC and SLC approaches resulting from the final round of the iterations were associated with the corresponding items to be used in the operational testing situations. The exposure control parameters of both SHC and SLC strategies utilized here were those derived using the conditional samples of 10,000 examinees. In addition to the above three methods for which the exposure control parameters were utilized to directly limit the overexposure of items, the procedure with no control over item exposure was also applied to the operational CAT administrations for comparison purposes.

While the NO procedure was included in the tables or figures under the desired rate label of .10 or .20, it only served to facilitate comparisons with the other procedures under these two desired maximum conditions. The maximum desired rates of .10 and .20 did not affect the performance of the NO procedure due to no utilization of the exposure control parameters. The differences between the results reported under the desired rates of .10 and .20 were simply due to sampling errors.

Described below are the results of all four methods under the various conditions with respect to each of these criteria: the item security, the test-retest and the peer-to-peer test overlap rates, the utilization of the item pool, and the CSEMs. The evaluation process was based on general observations

of the results, and the results of using the four item pools, as well as the results of specifying the desired maximum exposure rate as .10 or .20.

I. Item Security

The degrees of item security achieved by utilizing the various exposure control methods with regard to the entire examinee group are presented. In addition, the results were also analyzed in a conditional fashion by applying the exposure control methods to the examinees of a particular ability level. The maximum exposure rates conditionally observed at each ability level are plotted to show test security at the respective ability levels.

General Observations

Table 1 reports the summary statistics of the observed exposure rates as a result of applying the various exposure control methods to the entire examinee group. The N column lists the numbers of items that were administered to examinees at least once. Based on these items, the summary statistics were obtained. The column of the maximum exposure rate shows the degree of item security achieved by the various procedures in light of the examinee group as a whole.

For the NO procedure, the high values of the observed exposure rates were simply the consequences of the utilization of maximum item information selection with no exposure control. For the DP and SHC procedures, the maximum observed exposure rates were as low as the desired values. An exception was observed when the SHC method was used under the 360-item pool with the desired exposure rate of .10 where the high maximum was likely a result of reconsidering the selected but not administered items for an examinee's CAT. For the SLC procedure, the observed maximum exposure rates were as low as the desired value when it was specified to be .20; however, the observed maximums were higher than the desired rate when it was targeted at .10, except for the largest pool of 720 items. These results were fairly unsatisfactory but were not unexpected given that the maximum observed exposure rates were stabilized but to values higher than the desired rates in the derivation of the SLC exposure control parameters. It might be that the SLC procedure selected the next item from a list of items that did not include all available items in the pool.

The conditional maximum observed exposure rates at each ability level are shown in Figure 10. As displayed in these plots, when the item exposure rates were not controlled by the NO method, the conditionally observed maximums reached as high as 1.0. The conditional maximum observed exposure rates for the DP procedure were near .30 and .40 for the target exposure rates of .10 and .20 respectively. The values were higher at the two extreme ability levels, especially under the desired rate of .10 using the 360- and 480-item pools.

The results of employing both conditional procedures were the most satisfactory. The conditional maximum observed exposure rates appeared to be fairly stable over most of the ability continuum. The values were only slightly above the desired levels for the middle ability points, which accounted for most of the examinee population. The maximums were observed somewhat higher at both extremes due to insufficient numbers of items appropriate for administration at these extreme ability levels. Also, the conditional maximums were in general lower for the SHC procedure than for the SLC method, with greater differences at the two ends of the ability scale than at the middle. The reconsideration of selected but not administered items under the 360-item pool and the desired exposure rate of .10 might contribute in part to the high conditional observed exposure rates at the high end of the ability scale for the SHC algorithm.

The Effect of Item Pool Size

As shown in Table 1, when the item selection was not controlled for exposure rates, the maximum observed rate reached 1.0 for any item pool. This finding suggests that a large pool itself was not sufficient to guarantee test security. When the selection was based on the DP or SHC methodology, the employment of a larger pool did not lead the maximum observed exposure rates to drop. For the SLC procedure, as the pool size was increased, the maximum observed exposure rates were decreased under the desired exposure rate of .10. Given the unsatisfactory results in achieving the test security that the SLC algorithm performed, allowing more items of similar quality in the pool seemed to remedy the situation to some extent.

As the pool size was enlarged, the average exposure rate decreased for each method (see Table 1). These results were expected since there were more items of similar quality in larger pools to allow more choices for item selection for administration. The average observed exposure rates were reduced as larger pools were employed for the DP, SHC and SLC procedures. The extent to which the values were reduced for the NO procedure was small. It seems that the procedures with statistical control over item exposure were more likely to benefit from enlarging the pool size than with no statistical control in lowering the average observed exposure rates.

As displayed in Figure 10 under the desired rate of .10, when the pool size was expanded from of 360 to 480, the conditional maximum observed rates of the DP method were increased at the middle ability points. While under the desired rate of .20, the values were similar at the higher end but were increased below the ability point of .40. These results of higher observed conditional maximum rates under larger item pools were similar to Chang (1998), in that there might exist some items in larger pools which were especially appropriate for administration to cause the higher appearance frequencies in the pools.

When the pool size was increased from of 480 to 600, the conditional maximum observed exposure rates of the DP procedure were increased under the desired rate of .10 at the lower end of the ability scale, although the values decreased under the desired rate of .20, especially at the higher end. As the pool was enlarged to be of 720 items, there existed virtually no differences of the conditional maximum observed rates for the DP procedure for either of the two desired rates. The utilization of larger pools only profited this procedure in achieving higher test security under some conditions.

For the SHC method, increasing the pool size seemed not to lower the conditional maximum observed exposure rates, except for increasing the pool size from of 360 to 480 under the desired rate of .10 where the conditional maximum observed rates were substantially reduced. As to the SLC procedure, enlarging the pool size from of 360 to 480 and also from of 480 to 600 reduced the conditional maximum observed rates when the desired rate of .10 was employed. While under the desired rate of .20, when the pool size was increased from of 360 to 480, the conditional maximum values of both SHC and SLC procedures were observed to be lower for the middle but slightly higher for the lower end of the ability scale.

The Effect of the Desired Maximum Exposure Rate

For the target probability of .10, the DP and SHC procedures yielded very similar maximum observed exposure rates, which were all at the pre-specified rate (see Table 1). An exception was observed under the 360-item pool where the SHC method failed to control the exposure rates of the items to the desired value. The SLC procedure did not control the item exposure rates to the desired value of .10. Especially with the 360-item pool, the maximum observed exposure rates were as high as .17. As the target probability was loosened to .20, all approaches successfully limited the exposure rates of items to the desired value of .20.

II. Item Overlap

The performance of each method in terms of the test-retest overlap rates and the peer-to-peer overlap rates is evaluated below.

1. The Test-Retest Overlap Rate

The average values of the test-retest overlap rates (i.e., the test-retest mean overlap rates) are reported for each ability point.

General Observations

Figure 11 demonstrates that for each method, the overall patterns of the test-retest mean overlap rates were similar across the various conditions. The test-retest mean overlap rates of the NO procedure were large, especially higher at the very high ends. The DP, SHC and SLC methods produced

substantially smaller test-retest mean overlap rates, which were fairly stable across the entire ability continuum. Under the desired rate of .10, both conditional methods of SHC and SLC resulted in smaller test-retest mean overlap rates than the DP procedure while under the desired rate of .20, the DP procedure led to higher test-retest mean overlap rates at the middle ability points but slightly lower overlap rates at the two ends of the continuum.

The Effect of Item Pool Size

As shown in Figure 11, there was no substantial reduction in the test-retest mean overlap rates for any method when the pool size was increased. A careful inspection reveals that increasing the pool size from of 360 to 480 caused the test-retest mean overlap rates of the NO method to drop at some ability levels. But the reduction was not found when the pool size was increased from of 480 to 600. However, increasing the pool size from of 600 to 720 caused the test-retest mean overlap rates for the NO method to increase between ability points of -.40 and .40. One possible explanation for this phenomenon might be similar to that observed in Figure 10 where some maximum observed exposure rates were higher under a larger pool condition than a smaller pool. There might exist in the 720-item pool some items especially appropriate for examinees at these particular ability levels to cause their higher appearance frequencies while no control was exercised for item exposure.

Under the desired rate of .10, enlarging the pool size from of 360 to 480 caused the SLC procedure to drop in the test-retest mean overlap rates. But when the pool continued to increase in size, it seems to yield no reduction in these types of overlap rates. For the DP and SHC procedures, there was not much decrease in the test-retest overlap rates at any ability level as the pool was enlarged to any extent. When the desired maximum exposure rate was relaxed to .20, enlarging the pool size resulted in no decrease in the test-retest mean overlap rates at all ability levels for all three procedures.

The Effect of the Desired Maximum Exposure Rate

The configurations in Figure 11 show that for the DP procedure, relaxing the desired maximum exposure rate from .10 to .20 caused the test-retest mean overlap rates at the middle ability levels (-1.2 to 1.2) to increase for all item pools. For the SHC an SLC approaches, the penalty of relaxing the desired exposure rate to .20 on the increase of the test-retest mean overlap rates was fairly consistent, but with slightly more increase at the two ends of the ability continuum. Under the desired rate of .10, the SHC procedure resulted in the most satisfactory test-retest mean overlap rates, followed by the SLC procedure and the DP procedure. Under the desired rate of .20, the three methods yielded very similar test-retest mean overlap rates, but with higher rates at the middle range for the DP approach.

2. The Peer-to-Peer Overlap Rate

The values of the peer-to-peer overlap rates were expected to be smaller than the test-retest overlap rates, because these values represent the overlap percentages of items administered to examinees of randomly dissimilar abilities rather than of the same ability so that the same items were not as likely to be administered.

General Observations

As presented in Table 2, as long as the selection procedure did not control for item exposure, the peer-to-peer overlap rates could reach a value as high as 1.0, no matter what size of the item pool was used. On average, the peer-to-peer overlap rates for the NO method ranged from .34 to .37, indicating that the adaptive tests for any two examinees of this group contained more than ten identical items on average. The employment of the DP, SHC and SLC techniques produced relatively low maximum peer-to-peer overlap rates. Their average peer-to-peer overlap rates were also seen to be small. The SHC and SLC procedures resulted in the smallest average values for most conditions.

The full distributions of the peer-to-peer overlap rates in Figure 12 show that for the NO procedure, these types of overlap rates were evenly distributed over almost the entire range. For the three procedures that implemented statistical control over item exposure, small overlap rates of .10 to .20 were observed for the majority of the distribution. Their distributions were very similar under the desired rate of .20 but were somewhat different between the DP procedure and the two conditional procedures under the desired rate of .10.

The Effect of Item Pool Size

For the NO procedure, the greatest reduction in the average peer-to-peer overlap percentages was seen when the pool size was enlarged from of 360 to 600 (see Table 2). For the DP approach, enlarging the pool size caused the average peer-to-peer overlap rates to be reduced under the desired rate of .20, but led to no noticeable drop when the desired rate was limited to .10. For the SHC algorithm, successive reduction in these types of overlap rates was observed as the pool size was enlarged under both expected exposure rate conditions. Increasing the pool size seemed to lower the peer-to-peer overlap rates for the SLC procedure to a slightly greater extent than for the other methods, especially with the desired rate of .10.

The Effect of the Desired Maximum Exposure Rate

For the DP, SHC or SLC method, the relaxation of the desired maximum rate from .10 to .20 increased the average peer-to-peer overlap rates under any pool size condition. The extents to which the peer-to-peer overlap rates increased were small, but varied among these three procedures. It seems that

loosening the target exposure rate increased the average peer-to-peer overlap rates of the DP procedure by a slightly greater amount than the two conditional procedures.

III. Utilization of the Item Pool

The numbers and/or the percentages of items that were never administered are reported to indicate the extent to which the pool was utilized. The smaller the number of items relative to the whole pool that were never used (i.e., the unused percentage), the greater the utilization of the item pool. For each of the four procedures, the results of the pool utilization are presented with respect to the content categories, the item pool sizes, and the target maximum exposure rates.

General Observations

As presented in Table 3, under the NO procedure, more than half of the items in the smallest pool of 360 items had never been used and over 70% of the items in the largest pool of 720 items were never administered. These findings provide strong evidence that a large proportion of the item pool would be wasted if the exposure rates of items were not controlled.

While incorporating statistical methods of exposure control into the item selection process, noticeable reduction was observed in the percentages of items that were never used, particularly with the pre-specified rate of .10. For each of these strategies, every item in the 360-item pool was administered at least once under the desired rate of .10. The SHC and SLC procedures maintained complete or almost full utilization of the 480- and 600-item pools when the desired rate of .10 was specified. The DP procedure, in general, made good and stable use of the pool under the various conditions. As to the pool utilization with respect to the content categories, there were more unused items in the pre-algebra and plane geometry contents than in the other areas under many conditions for all three methods.

The Effect of Item Pool Size

Table 3 shows that while the item pool size was enlarged, all of the procedures resulted in a greater part of the item pool that had never been touched. That is, a larger pool was not utilized to an extent as great as that of a smaller pool for any procedure. These results were somewhat counter-intuitive since a larger pool contained items of similar quality to those in a smaller pool. Thus, it was expected that the percentages of unused items would be similar to those percentages in a smaller pool. However, recall the results with the item security and the test overlap rates where the higher maximum observed exposure rates and the higher test overlaps under some conditions occurred in larger pools rather than in smaller pools. These phenomena might be explained by the same reason given earlier; that there existed some items in larger pools especially appropriate for administration. Due to the frequent

administrations of such items, many other items were "ignored" so that the unused proportion was greater in larger pools than in smaller pools.

The Effect of the Desired Maximum Exposure Rate

As shown in Table 3, relaxing the desired exposure rate to .20 caused the unused percentages of the two conditional procedures to increase by a large extent. Such a huge increase in the unused percentages was not observed with the DP procedure. The increase in the unused percentages for the DP methodology was as small as 2% or 3% for the various item pools.

It can be noticed that when the expected rate was set to be .10, the employment of both conditional methods utilized the pool to a greater extent than the DP procedure except for the 360-item pool where all three methods achieved complete pool utilization. However, when the target exposure rate of .20 was used, the results were in the opposite direction--the employment of the DP methodology increased the pool utilization to a greater degree than that of the conditional procedures. Apparently, the extent to which the exposure rates were limited had different effects on the performance of the exposure control methods with respect to the pool utilization. Also, these differential effects varied under the various item pools.

IV. Conditional Standard Errors of Measurement

Due to the constraints on test content, measurement precision for the current CATs was expected to be compromised. Although the measurement precision might have been sacrificed to some extent in this study, examinees at any level of ability were ensured to receive CATs with appropriate content balance.

General Observations

The CSEM curves produced by using the various methods under the different pool sizes and the different expected exposure rates are displayed in Figure 13. The large conditional sample sizes of 3,000 examinees led to the smooth curves in the configurations. Because there were fewer items appropriate for administration for both extreme ability levels, the SEMs at the two ends were higher than those at the middle part of the ability scale. The effect of guessing caused higher SEMs at the lower end than at the upper end of the scale.

Figure 13 displays that, across all conditions, the DP, SHC and SLC strategies resulted in higher CSEMs than the NO procedure. These results were stronger at the two extremes than at the middle range. The price for better control of the item exposure rates with the DP, SHC and SLC algorithms was seen in the loss of measurement precision. This is, of course, not unexpected.

The Effect of Item Pool Size

A careful inspection of Figure 13 shows that when the control for item exposure was not exercised at all, increasing the pool size seems not to reduce the CSEMs for any ability level. For the three procedures that incorporated statistical control into the item selection process, the CSEMs were reduced to some extent as the pool size was increased from of 360 to 480, especially noticeable for the lower end of the ability scale. However, beyond the 480-item pool, the reduction in the CSEMs seems very small as two adjacent pool sizes were considered, especially under the desired exposure rate of .20. When the pool size was increased from of 360 to 600, the reduction in the CSEMs was more substantial, so the reduction would seem more noticeable for two pools of greater size differences. However, notice that under the desired rate of .20, the reduction in the CSEMs was not great when the pool size was increased from of 480 to 720. The findings tend to suggest that when the pool size is sufficiently large for the administration of items according to the CAT test length and the desired maximum exposure rate, measurement precision may not be further improved to a noticeable degree with the employment of a larger item pool. Utilizing a huge pool might not be necessary, since it does not increase the measurement precision in proportion.

The Effect of the Desired Maximum Exposure Rate

It can be seen in Figure 13 that the configurations under the two desired exposure rates were similar, but with greater differences in the SEMs among the various procedures under the desired rate of .10 than .20. When the desired rate was specified to be .10, the SHC algorithm produced the highest SEMs at any ability level for the four item pools. For the 360-item pool, its CSEMs were substantially higher than those of the other methods, especially at the extremes.

For the desired rate of .10, the high CSEM curves of the SHC method were followed by the SLC method for the various item pools. The SEMs produced by the DP procedure were lower than those of these two methods. The differences in the random errors for these three approaches were mitigated at the middle range of the ability continuum. Beyond this range, the loss of measurement precision due to the strict control of item exposure with the SHC and SLC methods appeared more prevalent.

In contrast to the results with the desired rate of .10 where the DP strategy produced lower SEMs at all ability levels than the SHC and SLC procedures, when the desired rate was loosened to .20, there existed no difference of SEMs among these three procedures for almost the entire ability continuum. These results indicate that when the desired exposure rate was as high as .20, both conditional methods would achieve the same measurement precision as the DP method.

Summary and Conclusions

Summary of Results

The current study examined the effectiveness of the SHC procedure, a modification of the SH method, in controlling the exposure rates of items in the CAT environment. Rather than deriving the exposure control parameters with respect to an entire examinee distribution representative of the real examinee population, the SHC approach derived the exposure control parameters in reference to a particular ability level. Its properties were compared with those of the DP and SLC procedures under four item pool sizes and two desired exposure rates. To produce stable estimates of the conditional exposure control parameters for both SHC and SLC methods while not consuming a large amount of time in the derivation process with large samples, this study attempted to determine appropriate conditional sample sizes for the development of exposure control parameters. The results of this study are summarized below.

The Effects of Conditional Sample Size on the Development of Exposure Control Parameters

The results indicated no effect of employing the four conditional sample sizes of 4,000, 6,000, 8,000 and 10,000 on the development of the exposure control parameters. The converging status of the SHC or SLC procedure seemed very similar when the various conditional sample sizes were utilized in carrying out the iterations. These findings suggested that employing a conditional sample size as small as 4,000 would be sufficient in producing similar estimates of the maximum observed exposure rates to employing a conditional sample size as large as of 10,000.

Advantages and Disadvantages of the Various Methods

The results yielded by the DP procedure were in general favorable. This procedure controlled the frequencies of item use as well as the overlap percentages across tests to a satisfactory degree. Also, it made good and consistent use of items in the pool. It was inevitable that such results led to increases in the errors of the ability estimates. However, in light of the test security, the percentage overlap and the degree of pool utilization this procedure achieved, the magnitudes of the CSEMs were judged tolerably small, except at the extreme ability levels.

The SHC method produced the most satisfactory results in terms of item security (both based on the entire examinee group and the examinees of a particular ability level) and the test overlap rates, followed by the SLC method. Except when the 600- and 720-item pools were employed in combination with the target exposure rate of .20 for which the degree of the pool usage was somewhat disappointing, the item pools were utilized to a great extent by using these two conditional procedures. However, the CSEMs were increased due to the strict control of the item exposure, especially at both extreme ability

levels. When developing the conditional exposure control parameters, the time consumed in the derivation process for the SHC procedure was less than that for the SLC procedure.

The Effects of Item Pool Size

The results showed that as long as the control for item exposure was not exercised, optimal items could be administered to almost every examinee under any of the four item pools. A large item pool itself, therefore, was not sufficient to guarantee test security. As to the DP, SHC and SLC procedures, the maximum observed exposure rates were not reduced when the pool size was increased, suggesting that the employment of larger pools containing items of similar quality to those of smaller pools did not profit the methods of exposure control in achieving higher test security. An exception was observed where the SLC procedure was employed under the target exposure rate of .10.

It was not surprising to see that while the pool size was enlarged, the average exposure rates of the items were reduced. However, the extent to which the average exposure rates decreased with the employment of larger pools was very small and varied among the methods. Enlarging the item pool size seemed more profitable for the three procedures that implemented statistical control over the item exposure than the NO procedure in reducing the average exposure rates.

The employment of larger pools slightly lowered both the test-retest and the peer-to-peer overlap rates for all four procedures. But, the impact was very small in general. Also, for all procedures, the item pool was even wasted to a greater extent when its size was enlarged. Regarding the CSEMs, the extent to which these errors decreased with larger pools varied among the methods. The findings seemed to suggest that when the pool size is sufficiently large for the administration of items according to the CAT test length and the desired maximum exposure rate, random errors in the ability estimates may not be further reduced with the employment of a larger item pool. Utilizing a huge pool might not be necessary, since it does not increase the measurement precision in proportion.

The Effects of Desired Maximum Exposure Rate

The results showed that the relaxation of the desired exposure rates from .10 to .20 differentially affected the performance of the exposure control methods in producing the maximum observed exposure rates. The performance of the procedures in controlling the test overlap rates and in utilizing the item pool was affected differently by the extent to which the exposure rates were limited. In terms of the errors of CSEMs, the results indicated that the effects of relaxing the desired exposure rates also varied among the procedures.

Conclusions

The effectiveness of the SHC method was investigated in the current study. The findings suggested that the performance of this approach was competitive to that of the SLC procedure. Although developing the exposure control parameters with respect to each ability level was very tedious, compared with the SLC method, the SHC procedure was more efficient in preparing these exposure parameters.

The issue of the appropriate conditional sample sizes in deriving the exposure control parameters was considered in this study. The results showed that employing a conditional sample size as large as of 10,000 at each ability level may not be necessary for producing stable estimates of the maximum observed exposure rates for the SHC or SLC method. A conditional sample as small as 4,000 examinees led to similar convergence of the conditional procedures as well. Smaller conditional sample sizes might still be attempted to further reduce the tedium in the iteration process.

The properties of the DP, SHC and SLC algorithms were explored and compared under the four item pool sizes and the two desired exposure rates with respect to item security, test-retest and peer-to-peer overlap rates, utilization of the item pool, and CSEMs. The better control of the item exposure rates and the test overlap rates, as well as better utilization of the item pool was accompanied by increases in the errors of the ability estimates. Particularly for examinees at the two ends of the ability scale, the increases in the errors were more substantial due to the fact that there were fewer items statistically adequate for administration for examinees at the extremes than at the middle.

The findings for the SHC and SLC procedures revealed that when an adaptive test of 30 items is delivered using a pool as small as of 360 items and a desired rate as strict as .10, optimal test security might not be ensured due to a lack of items needed for selection. When the exposure rate of .10 is desired, pools containing more than 360 items might be needed. In fact, it may be that this desired maximum exposure rate of .10 is unrealistically close to the average exposure rate of all items in the pool, as mathematically equal to the test length divided by the pool size (in this case, the average exposure rate is 30/360 = .083), so that this desired maximum rate cannot be achieved with the pool size of 360 and the test length of 30 (M. L. Stocking, personal communication, March 24, 2000). Stocking suggested that a larger pool or shorter test, or perhaps a pool with different properties may be employed to remedy this situation. Although a rule of thumb was suggested by Stocking (1994) that in order to support a fixed-length adaptive test of roughly one-half the length of the parallel linear form, it is necessary that the CAT item pool contains at least six to eight typical linear forms, it is also important to realize that a CAT item pool containing only six typical linear forms would not be large enough when the exposure rates of items are being strictly controlled by the conditional exposure control methods. Further studies are needed to provide more information about the appropriate item pool size in relation to a desired exposure rate.

Consistent with the findings in Chang (1998) and Davey and Parshall (1995), a large item pool itself did not appear sufficient to guarantee test security. Only by incorporating statistical mechanisms in the item selection procedure can the goal to improve test security be accomplished in the CAT environment. The results for the pool size effect indicated that the performance of the various exposure control methods was affected differentially by the size of the item pool. But, in general, when the pool size was sufficiently large for the CATs administrations according to the target exposure rate control, there seems not necessary to employ a huge item pool. Employing larger item pools did not necessarily better ensure test security or lower test overlap rates, nor necessarily increase the measurement precision. Instead, a large proportion of the pools could just be wasted. The issue of determining the optimal item pool size is complicated since it relates to many factors such as the specific features of the exposure control algorithms, the desired maximums of the exposure rates, the length of the CATs, and the properties and structures of the item pools.

Regarding the findings on the desired exposure rate, the effects of restricting the exposure rates at .10 on the performance of the methods were not consistent with those of restricting the exposure rates at .20. The extent to which the desired exposure rate affected the behavior of the various exposure control methods was subject to the specific features of the algorithms in combination with the size and structure of the item pools.

Since scale scores are reported for examinees taking the existing P&P ACT-Math, results of the current CAT study would carry more practical meaning if the score reporting metric had been adopted rather than the theta metric. Also, this study concerned only discrete types of items and took into account only the content matter of items during the item selection process. The behavior of the SHC algorithm as well as the DP and SLC procedures remains unknown under complex but more realistic adaptive testing situations, such as contexts having blocks of items associated with reading passages and other types of nonstatistical constraints such as the overlap or the item set constraints. Moreover, since the properties of the various exposure control approaches were evaluated within the scope of estimating examinees' abilities in the CAT environment, the effectiveness and psychometric features of the exposure control procedures within the certification or licensure CAT context cannot be adequately implied based on the results of this research. There is a clear need for further investigations.

In conclusion, among the three exposure control algorithms investigated in this study, the SHC procedure best served the purposes of controlling the observed exposure rates to the desired values as well as producing the lowest test overlap rates, followed by the SLC method. In terms of these two criteria, the DP procedure performed slightly less well; but, the DP method utilized the item pools to a satisfactory extent overall. However, the trade-offs accompanying such desirable outcomes of these three procedures were that their CSEMs of the ability estimates were increased to some extent, particularly at

both extreme ability levels. This poses a dilemma for testing programs, and the choice of procedures would depend on the purposes and needs of administering the adaptive tests, as well as on the ways the tests are implemented. In general, within the continuous testing environment of the CATs where test security might be at great risk, imposing control for the overexposure of optimal items might be worth the sacrifice of some measurement precision.

Given the increasing interest in and use of CATs, the results of this study have improved on the guidelines for psychometric researchers and test practitioners to select most appropriate procedures to control item exposure. The findings in this study have also provided useful insights on how the item pool sizes and how the maximum item exposure rates affect the performance of the exposure control methods. With better control of item exposure in advance of testing, test security concerns in the CAT environment should be lessened.

References

ACT. (1997). *ACT assessment technical manual*. Iowa City, IA: ACT, Inc.

Chang, S. W. (1998). *A comparative study of item exposure control methods in computerized adaptive testing*. Unpublished doctoral dissertation, The University of Iowa. Iowa City, IA.

Davey, T., & Parshall, C. G. (1995, April). *New algorithms for item selection and exposure control with computerized adaptive testing*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Featherman, C. M., Subhiyah, R. G., & Hadadi, A. (1996, April). *Effects of randomesque item selection on CAT item exposure rates and proficiency estimation under the 1- and 2-PL models*. Paper presented at the annual meeting of the American Educational Research Association, New York.

Hetter, R. D., & Sympson, J. B. (1997). Item exposure control in CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 141-144). Washington, DC: American Psychological Association.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2*(4), 359-375.

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), *New horizons in testing : Latent trait test theory and computerized adaptive testing* (pp. 223-236). New York: Academic Press.

Mislevy, R. J., & Bock, R. D. (1990). Item analysis and test scoring with binary logistic models. *BILOG 3*. Chicago, IL: Scientific Software, Inc.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70*(350), 351-356.

Parshall, C. G., Davey, T., & Nering, M. L. (1998, April). *Test development exposure control for adaptive testing*. Paper presented at the annual meeting the National Council on Measurement in Education, San Diego.

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 35*(4). 311-327.

Stocking, M. L. (1994). *Three practical issues for modern adaptive testing item pools* (Research Report 94-5). Princeton, NJ: Educational Testing Service.

Stocking, M. L., & Lewis, C. (1995a). *A new method of controlling item exposure in computerized adaptive testing* (Research Report 95-25). Princeton, NJ: Educational Testing Service.

Stocking, M. L., & Lewis, C. (1995b). *Controlling item exposure conditional on ability in computerized adaptive testing* (Research Report 95-24). Princeton, NJ: Educational Testing Service.

Stocking, M. L., & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 23*(1), 57-75.

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17*(3), 277-292.

Swanson, L., & Stocking, M. L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement, 17*(2), 151-166.

Sympson, J. B., & Hetter, R. D. (1985, October). *Controlling item-exposure rates in computerized adaptive testing.* Paper presented at the 17th annual meeting of the Military Testing Association. San Diego, CA: Navy Personnel Research and Development Center.

Table 1. Results of Observed Exposure Rates for the Various Methods by Item Pool and Desired Exposure Rate

r = .10

	Method	N	Mean	SD	Minimum	Maximum
The 360-Item Pool	NO	153	0.19608	0.18574	0.00004	1.00000
	DP	360	0.08333	0.02912	0.00134	0.10754
	SHC	360	0.08333	0.02429	0.00674	0.16318
	SLC	360	0.08333	0.04404	0.00240	0.16806
The 480-Item Pool	NO	171	0.17544	0.17542	0.00004	1.00000
	DP	470	0.06383	0.03934	0.00002	0.10844
	SHC	480	0.06250	0.02968	0.00122	0.10154
	SLC	476	0.06303	0.03643	0.00004	0.13512
The 600-Item Pool	NO	172	0.17442	0.16947	0.00002	1.00000
	DP	556	0.05396	0.03984	0.00002	0.10702
	SHC	599	0.05008	0.02900	0.00012	0.10078
	SLC	586	0.05119	0.03121	0.00002	0.11464
The 720-Item Pool	NO	186	0.16129	0.17059	0.00004	1.00000
	DP	623	0.04815	0.03981	0.00002	0.10758
	SHC	706	0.04249	0.02770	0.00002	0.09962
	SLC	684	0.04386	0.02807	0.00002	0.10226

r = .20

	Method	N	Mean	SD	Minimum	Maximum
The 360-Item Pool	NO	153	0.19608	0.18574	0.00004	1.00000
	DP	359	0.08357	0.06896	0.00004	0.20958
	SHC	353	0.08499	0.05635	0.00004	0.19972
	SLC	347	0.08646	0.05746	0.00002	0.19862
The 480-Item Pool	NO	171	0.17544	0.17542	0.00004	1.00000
	DP	457	0.06565	0.06542	0.00002	0.20920
	SHC	438	0.06849	0.05168	0.00002	0.19734
	SLC	433	0.06928	0.05111	0.00002	0.19822
The 600-Item Pool	NO	172	0.17442	0.16947	0.00002	1.00000
	DP	546	0.05495	0.05975	0.00002	0.21022
	SHC	498	0.06024	0.04783	0.00002	0.20318
	SLC	489	0.06135	0.04696	0.00002	0.20022
The 720-Item Pool	NO	186	0.16129	0.17059	0.00004	1.00000
	DP	607	0.04942	0.05677	0.00002	0.20744
	SHC	535	0.05607	0.04528	0.00002	0.19746
	SLC	525	0.05714	0.04424	0.00002	0.19972

Note. The descriptive statistics were based on items that were used at least once.

BEST COPY AVAILABLE

Table 2. The Peer-to-Peer Test Overlap Rates for the Various Methods by Item Pool and Desired Exposure Rate

			r = .10		
	Method	Mean	SD	Minimum	Maximum
The 360-Item Pool	NO	0.37161	0.28811	0.03333	1.00000
	DP	0.09337	0.06457	0.00000	0.43333
	SHC	0.09018	0.05059	0.00000	0.33333
	SLC	0.10714	0.06295	0.00000	0.40000
The 480-Item Pool	NO	0.35067	0.28382	0.03333	1.00000
	DP	0.08845	0.07098	0.00000	0.43333
	SHC	0.07669	0.04969	0.00000	0.36667
	SLC	0.08413	0.05772	0.00000	0.36667
The 600-Item Pool	NO	0.33873	0.28354	0.03333	1.00000
	DP	0.08339	0.07247	0.00000	0.43333
	SHC	0.06682	0.05002	0.00000	0.30000
	SLC	0.07018	0.05479	0.00000	0.40000
The 720-Item Pool	NO	0.34187	0.28260	0.03333	1.00000
	DP	0.08118	0.07379	0.00000	0.46667
	SHC	0.06072	0.05066	0.00000	0.46667
	SLC	0.06170	0.05256	0.00000	0.33333

			r = .20		
	Method	Mean	SD	Minimum	Maximum
The 360-Item Pool	NO	0.37161	0.28811	0.03333	1.00000
	DP	0.14047	0.08833	0.00000	0.53333
	SHC	0.12235	0.07764	0.00000	0.43333
	SLC	0.12443	0.08195	0.00000	0.50000
The 480-Item Pool	NO	0.35067	0.28382	0.03333	1.00000
	DP	0.13027	0.09189	0.00000	0.56667
	SHC	0.10817	0.08009	0.00000	0.46667
	SLC	0.10654	0.08014	0.00000	0.46667
The 600-Item Pool	NO	0.33873	0.28354	0.03333	1.00000
	DP	0.11970	0.09129	0.00000	0.53333
	SHC	0.09813	0.08064	0.00000	0.46667
	SLC	0.09675	0.07999	0.00000	0.46667
The 720-Item Pool	NO	0.34187	0.28260	0.03333	1.00000
	DP	0.11517	0.09076	0.00000	0.53333
	SHC	0.09230	0.08014	0.00000	0.46667
	SLC	0.09137	0.07972	0.00000	0.50000

Table 3. Numbers of Items Never Used for the Various Methods by Content Category

r = .10

The 360-Item Pool

Content Category	NO	DP	SHC	SLC
Pre-Algebra	50 (60%)	0 (0%)	0 (0%)	0 (0%)
Elementary Algebra	35 (58%)	0 (0%)	0 (0%)	0 (0%)
Intermediate Algebra	33 (61%)	0 (0%)	0 (0%)	0 (0%)
Coordinate Geometry	28 (52%)	0 (0%)	0 (0%)	0 (0%)
Plane Geometry	47 (56%)	0 (0%)	0 (0%)	0 (0%)
Trigonometry	14 (58%)	0 (0%)	0 (0%)	0 (0%)
Total Unused	207 (58%)	0 (0%)	0 (0%)	0 (0%)

The 480-Item Pool

Content Category	NO	DP	SHC	SLC
Pre-Algebra	77 (69%)	5 (4%)	0 (0%)	3 (3%)
Elementary Algebra	49 (61%)	0 (0%)	0 (0%)	0 (0%)
Intermediate Algebra	50 (69%)	1 (1%)	0 (0%)	0 (0%)
Coordinate Geometry	42 (58%)	0 (0%)	0 (0%)	0 (0%)
Plane Geometry	71 (63%)	4 (4%)	0 (0%)	1 (1%)
Trigonometry	20 (63%)	0 (0%)	0 (0%)	0 (0%)
Total Unused	309 (64%)	10 (2%)	0 (0%)	4 (1%)

The 600-Item Pool

Content Category	NO	DP	SHC	SLC
Pre-Algebra	103 (74%)	20 (14%)	1 (1%)	10 (7%)
Elementary Algebra	69 (69%)	4 (4%)	0 (0%)	0 (0%)
Intermediate Algebra	65 (72%)	2 (2%)	0 (0%)	1 (1%)
Coordinate Geometry	61 (68%)	3 (3%)	0 (0%)	0 (0%)
Plane Geometry	101 (72%)	15 (11%)	0 (0%)	3 (2%)
Trigonometry	29 (73%)	0 (0%)	0 (0%)	0 (0%)
Total Unused	428 (71%)	44 (7%)	1 (0%)	14 (2%)

The 720-Item Pool

Content Category	NO	DP	SHC	SLC
Pre-Algebra	129 (77%)	39 (23%)	10 (6%)	21 (13%)
Elementary Algebra	88 (73%)	13 (11%)	0 (0%)	3 (2%)
Intermediate Algebra	79 (73%)	6 (6%)	1 (1%)	1 (1%)
Coordinate Geometry	74 (69%)	7 (6%)	0 (0%)	1 (1%)
Plane Geometry	128 (76%)	31 (18%)	3 (2%)	9 (5%)
Trigonometry	36 (75%)	1 (2%)	0 (0%)	1 (2%)
Total Unused	534 (74%)	97 (13%)	14 (2%)	36 (5%)

Table 3. (Continued)

r = .20

The 360-Item Pool

Content Category	NO	DP	SHC	SLC
Pre-Algebra	50 (60%)	1 (1%)	3 (4%)	8 (10%)
Elementary Algebra	35 (58%)	0 (0%)	0 (0%)	0 (0%)
Intermediate Algebra	33 (61%)	0 (0%)	1 (2%)	1 (2%)
Coordinate Geometry	28 (52%)	0 (0%)	0 (0%)	0 (0%)
Plane Geometry	47 (56%)	0 (0%)	3 (4%)	4 (5%)
Trigonometry	14 (58%)	0 (0%)	0 (0%)	0 (0%)
Total Unused	207 (58%)	1 (0%)	7 (2%)	13 (4%)

The 480-Item Pool

Content Category	NO	DP	SHC	SLC
Pre-Algebra	77 (69%)	13 (12%)	20 (18%)	21 (19%)
Elementary Algebra	49 (61%)	0 (0%)	3 (4%)	5 (6%)
Intermediate Algebra	50 (69%)	0 (0%)	3 (4%)	3 (4%)
Coordinate Geometry	42 (58%)	3 (4%)	3 (4%)	3 (4%)
Plane Geometry	71 (63%)	7 (6%)	12 (11%)	14 (13%)
Trigonometry	20 (63%)	0 (0%)	1 (3%)	1 (3%)
Total Unused	309 (64%)	23 (5%)	42 (9%)	47 (10%)

The 600-Item Pool

Content Category	NO	DP	SHC	SLC
Pre-Algebra	103 (74%)	27 (19%)	36 (26%)	39 (28%)
Elementary Algebra	69 (69%)	5 (5%)	15 (15%)	15 (15%)
Intermediate Algebra	65 (72%)	2 (2%)	13 (14%)	13 (14%)
Coordinate Geometry	61 (68%)	3 (3%)	6 (7%)	8 (9%)
Plane Geometry	101 (72%)	17 (12%)	28 (20%)	31 (22%)
Trigonometry	29 (73%)	0 (0%)	4 (10%)	5 (13%)
Total Unused	428 (71%)	54 (9%)	102 (17%)	111 (18%)

The 720-Item Pool

Content Category	NO	DP	SHC	SLC
Pre-Algebra	129 (77%)	44 (26%)	55 (33%)	55 (33%)
Elementary Algebra	88 (73%)	13 (11%)	26 (22%)	27 (23%)
Intermediate Algebra	79 (73%)	8 (7%)	26 (24%)	29 (27%)
Coordinate Geometry	74 (69%)	8 (7%)	18 (17%)	19 (18%)
Plane Geometry	128 (76%)	38 (23%)	49 (29%)	54 (32%)
Trigonometry	36 (75%)	2 (4%)	11 (23%)	11 (23%)
Total Unused	534 (74%)	113 (16%)	185 (26%)	195 (27%)

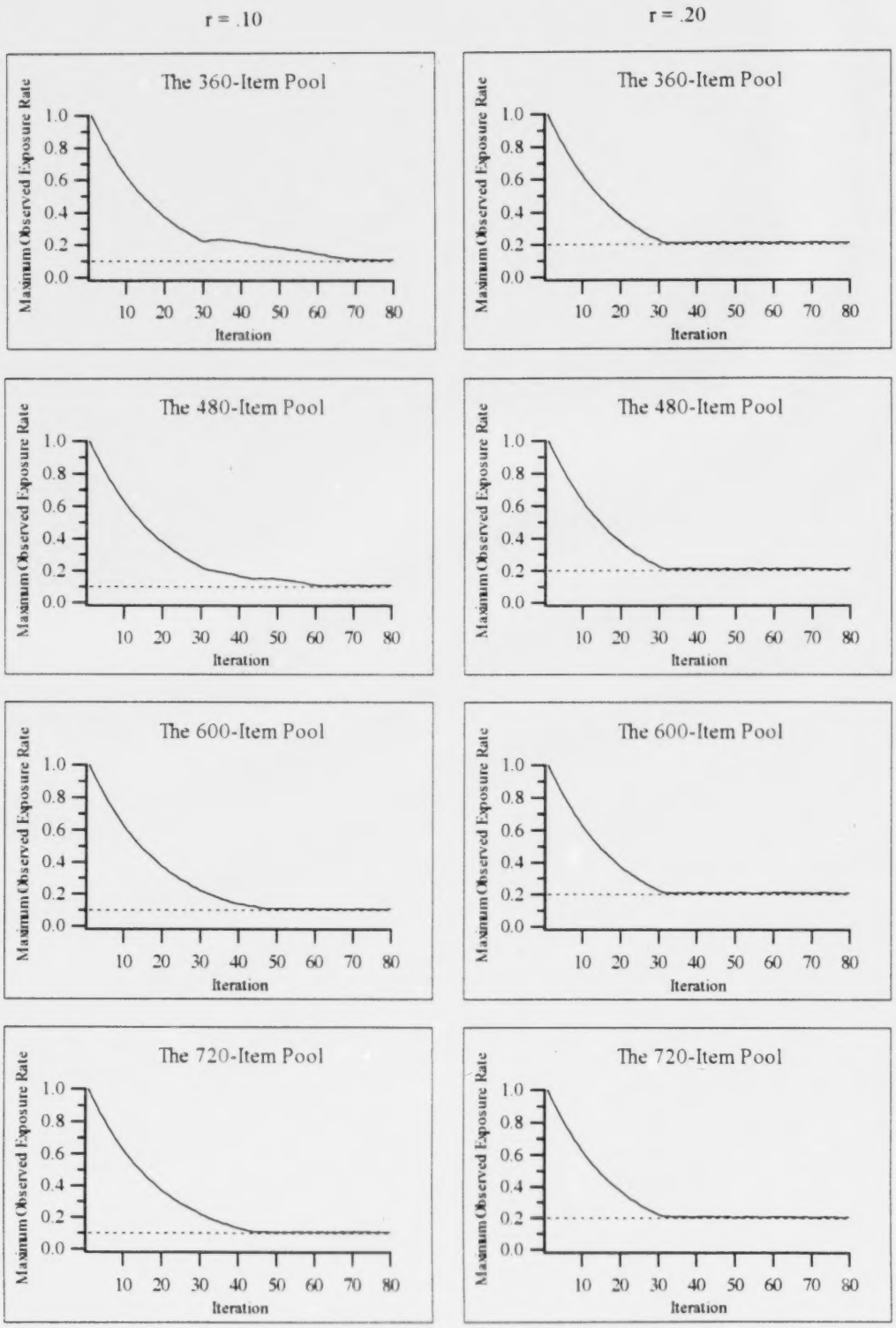

Figure 1. Iterations of the DP Procedure for the Various Item Pools

BEST COPY AVAILABLE

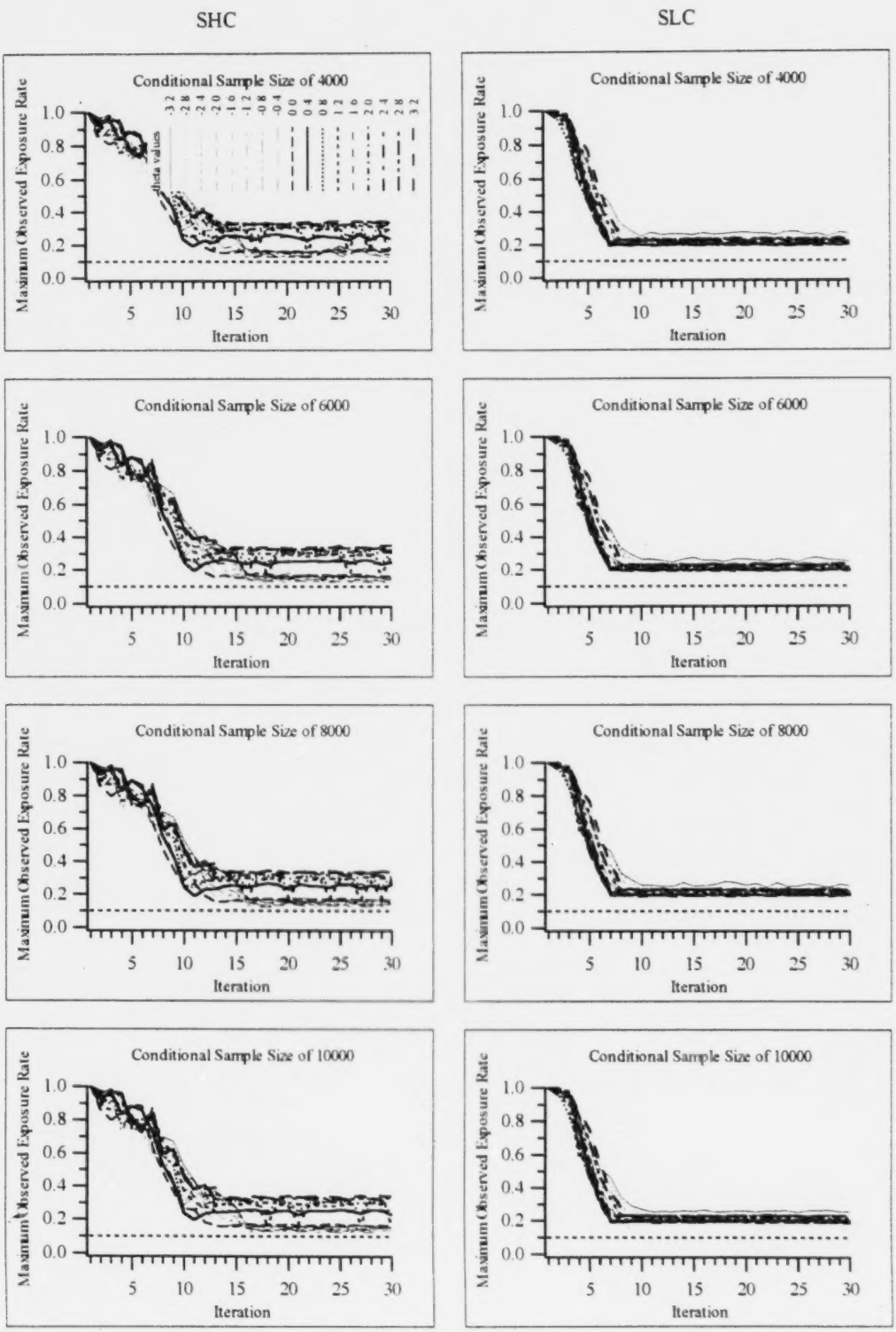

Figure 2. Iterations for the Various Conditional Sample Sizes with the 360-Item Pool and r = .10

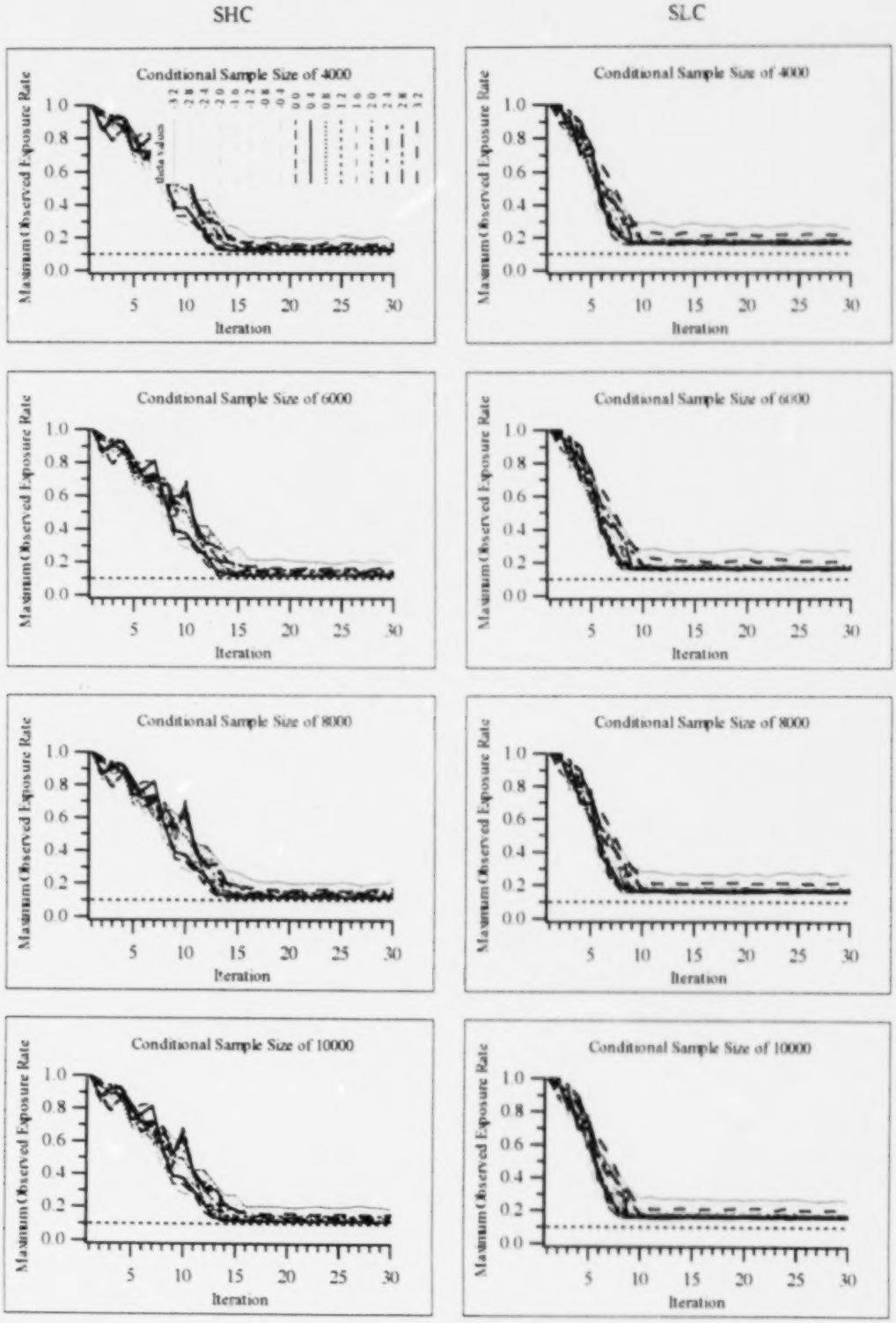

Figure 3. Iterations for the Various Conditional Sample Sizes with the 480-Item Pool and r = .10

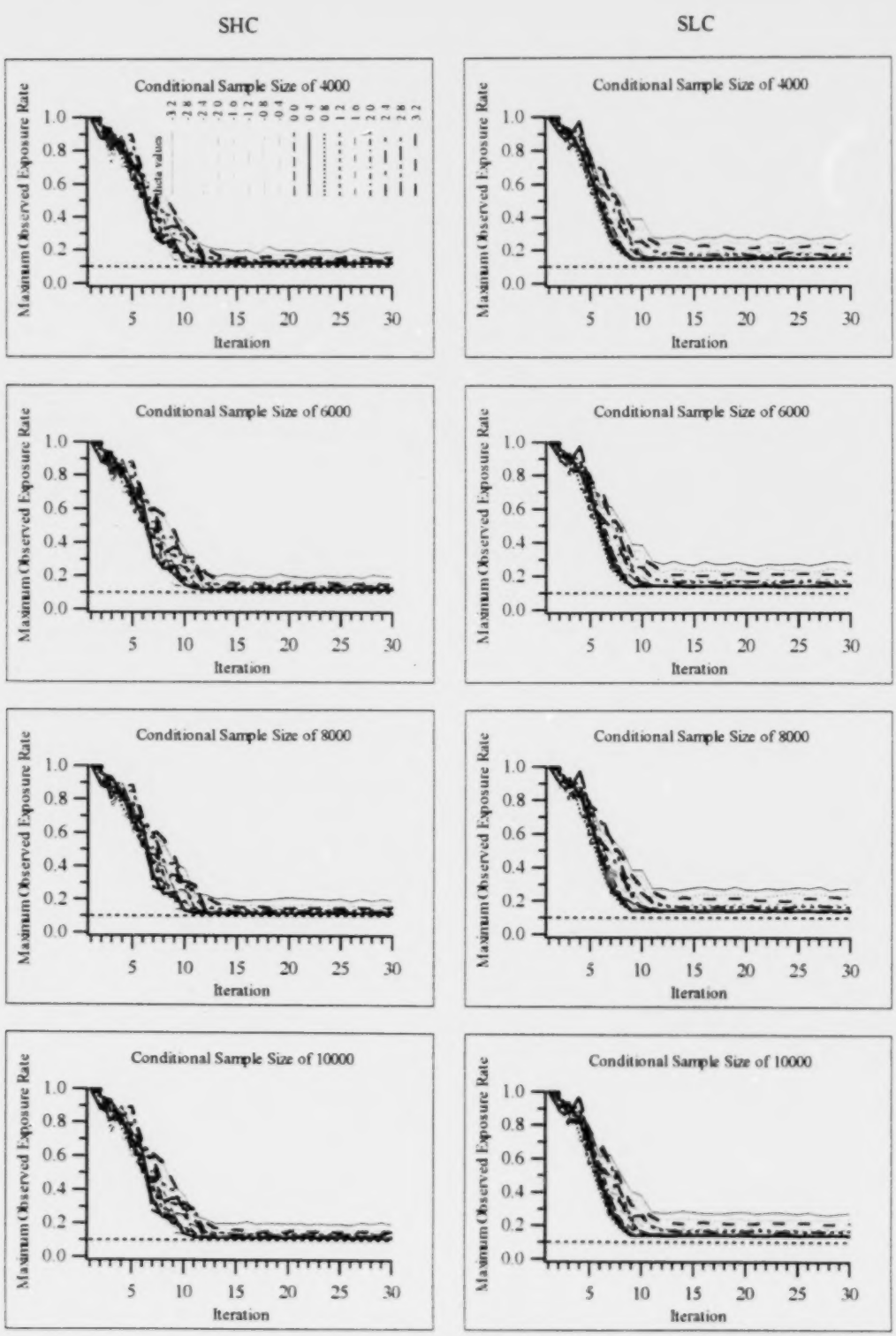

Figure 4. Iterations for the Various Conditional Sample Sizes with the 600-Item Pool and r = .10

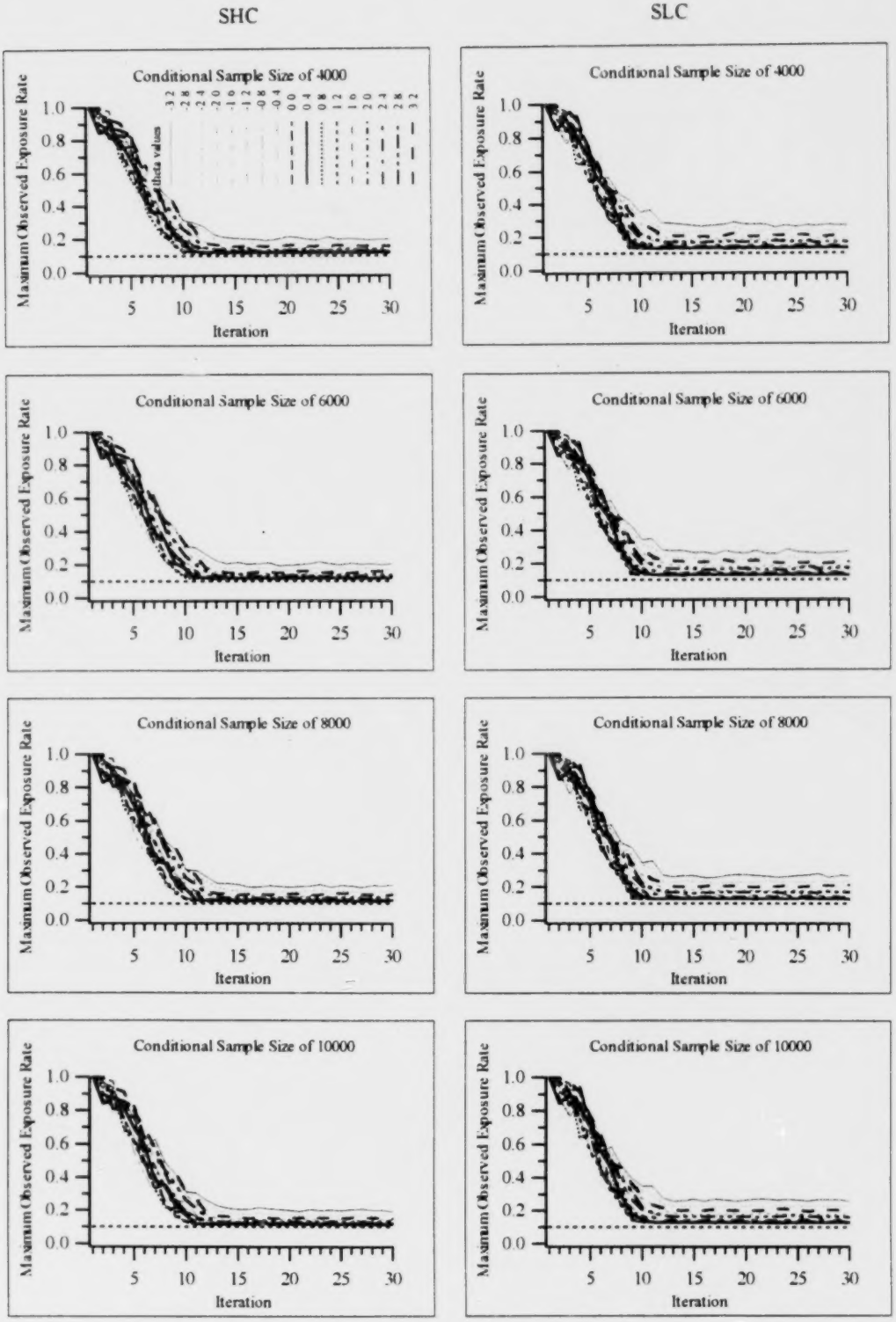

Figure 5. Iterations for the Various Conditional Sample Sizes with the 720-Item Pool and r = .10

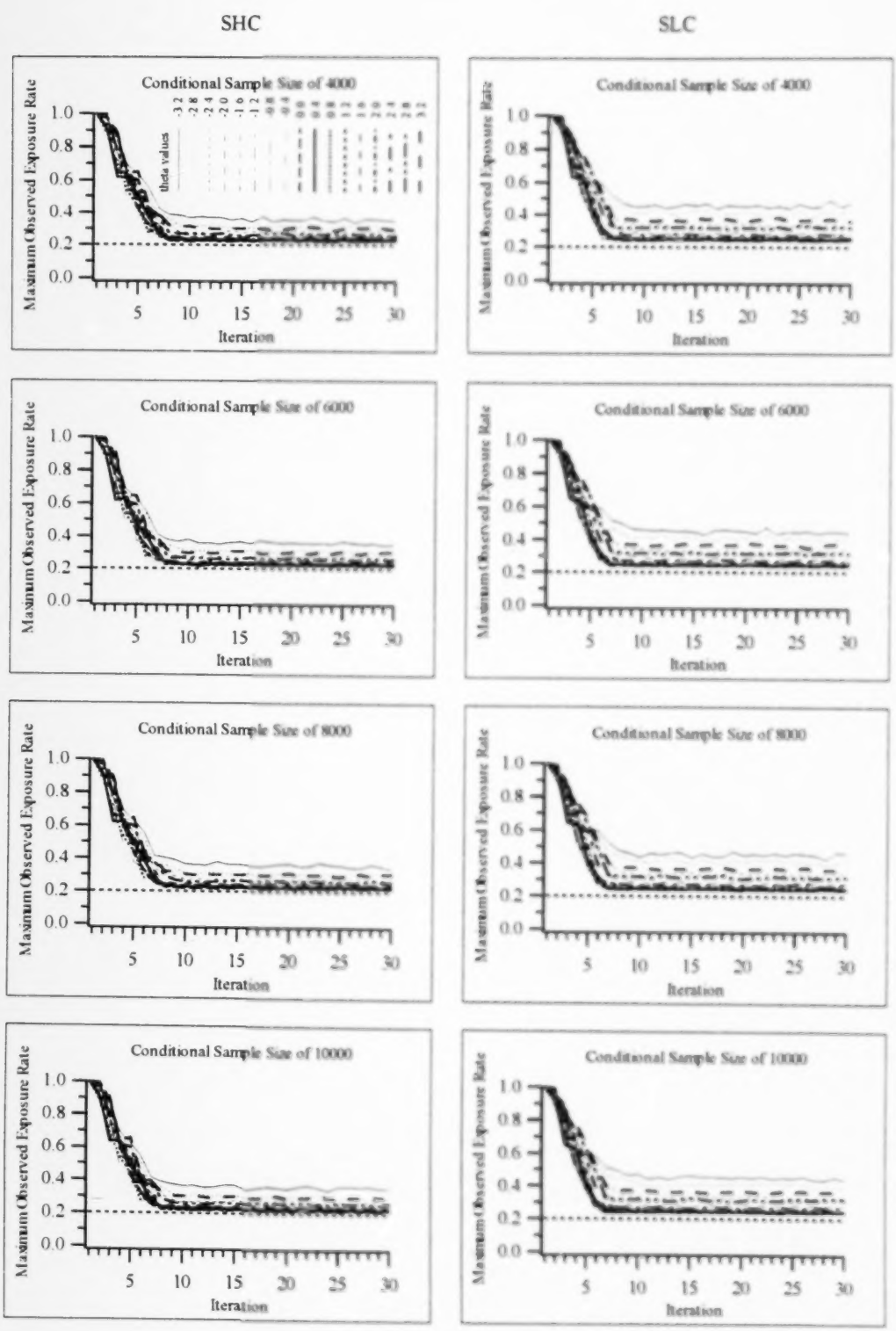

Figure 6. Iterations for the Various Conditional Sample Sizes with the 360-Item Pool and r = .20

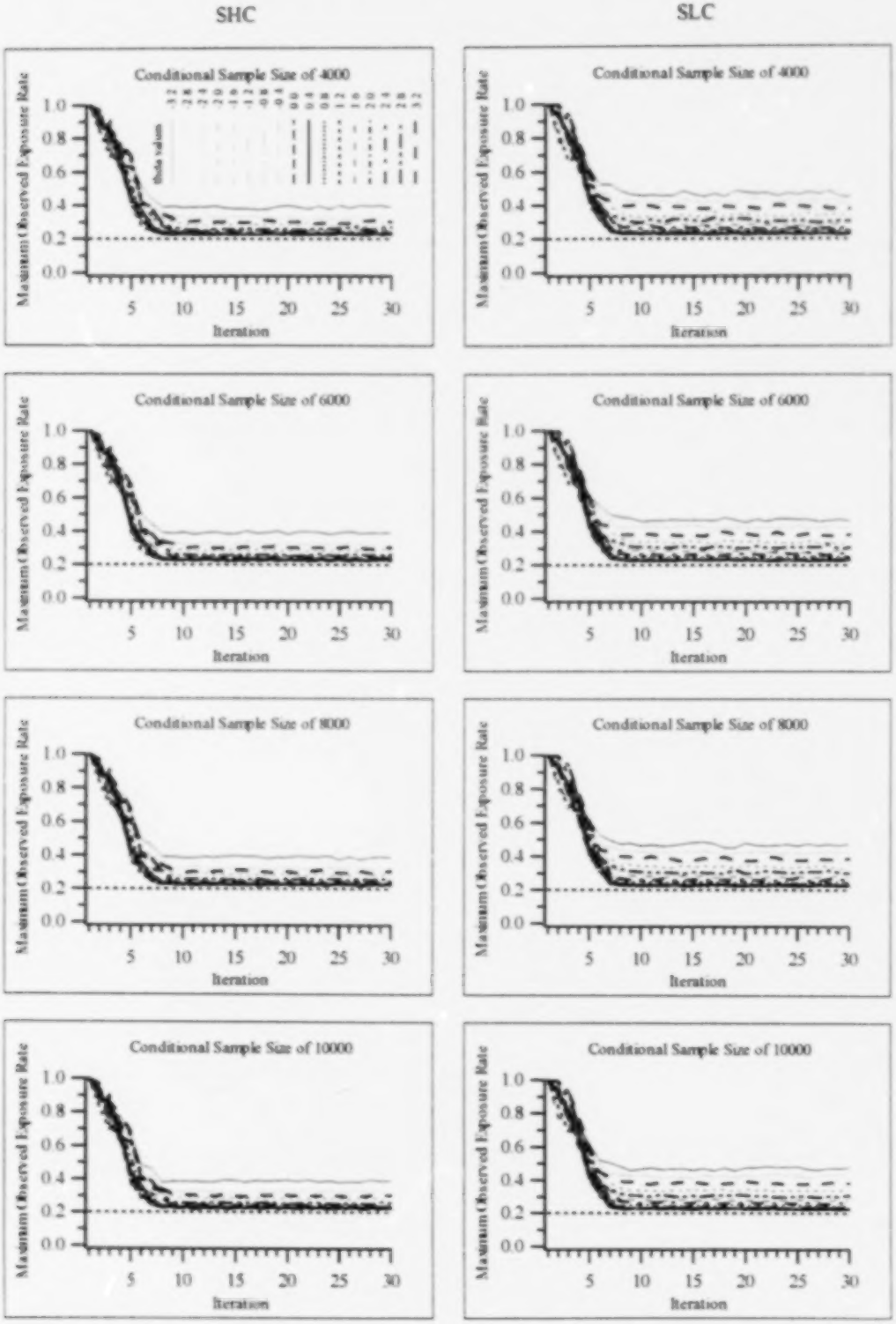

Figure 7. Iterations for the Various Conditional Sample Sizes with the 480-Item Pool and r = .20

BEST COPY AVAILABLE

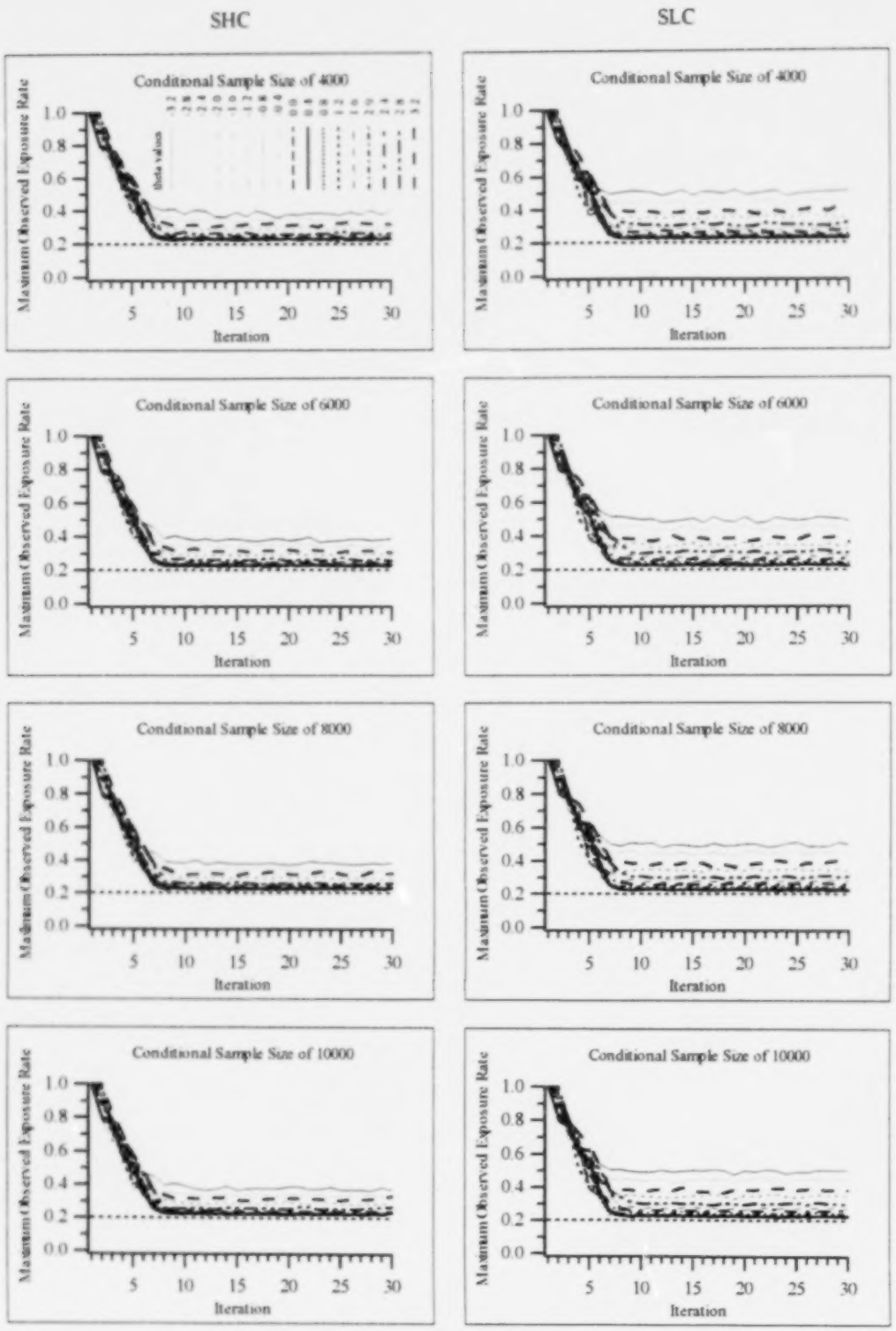

Figure 8. Iterations for the Various Conditional Sample Sizes with the 600-Item Pool and r = .20

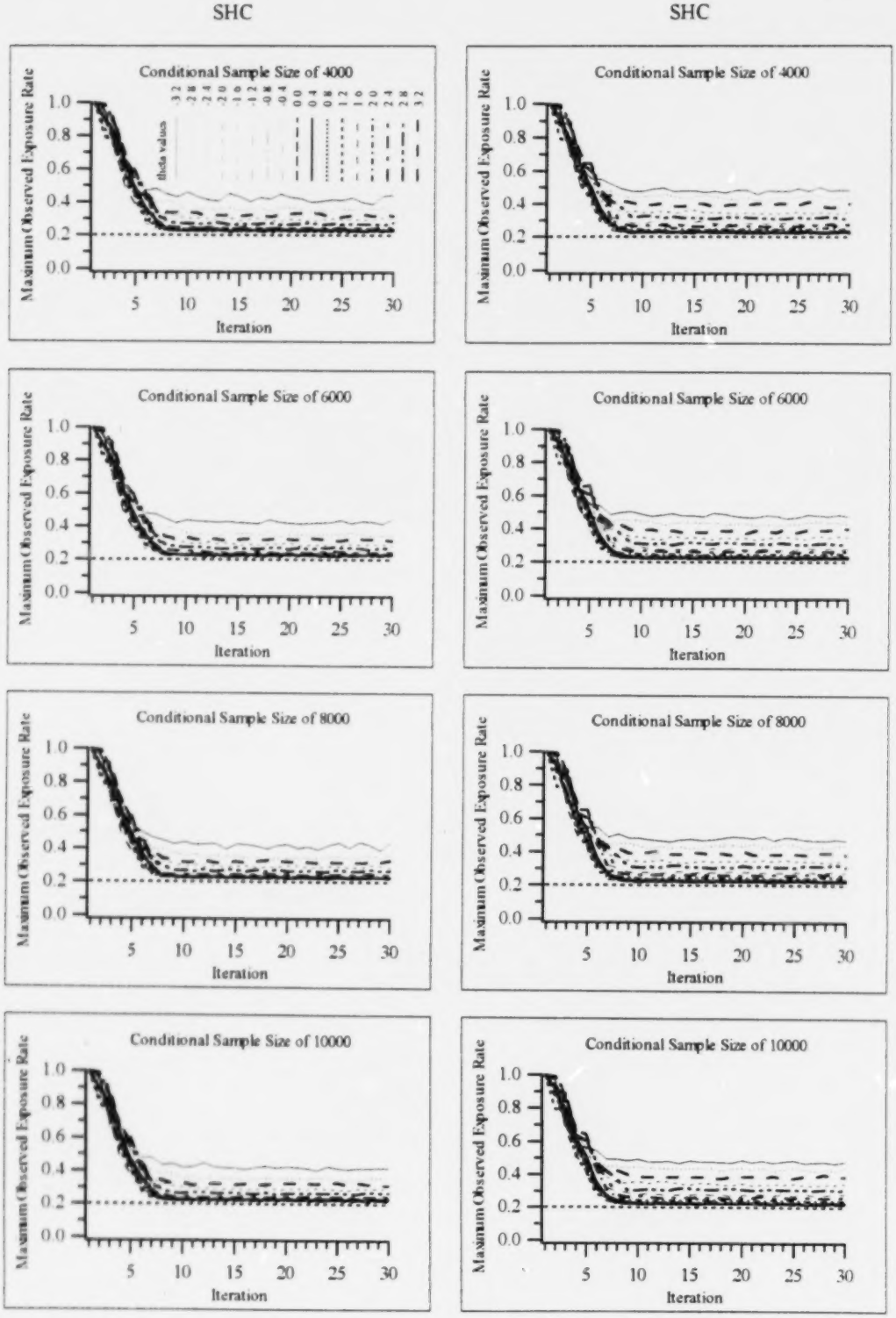
SHC
SHC
Conditional Sample Size of 4000
Conditional Sample Size of 4000
Conditional Sample Size of 6000
Conditional Sample Size of 6000
Conditional Sample Size of 8000
Conditional Sample Size of 8000
Conditional Sample Size of 10000
Conditional Sample Size of 10000
Maximum Observed Exposure Rate
Iteration
theta values
-3.2 -2.8 -2.4 -2.0 -1.6 -1.2 -0.8 -0.4 0.0 0.4 0.8 1.2 1.6 2.0 2.4 2.8 3.2

Figure 9. Iterations for the Various Conditional Sample Sizes with the 720-Item Pool and r = .20

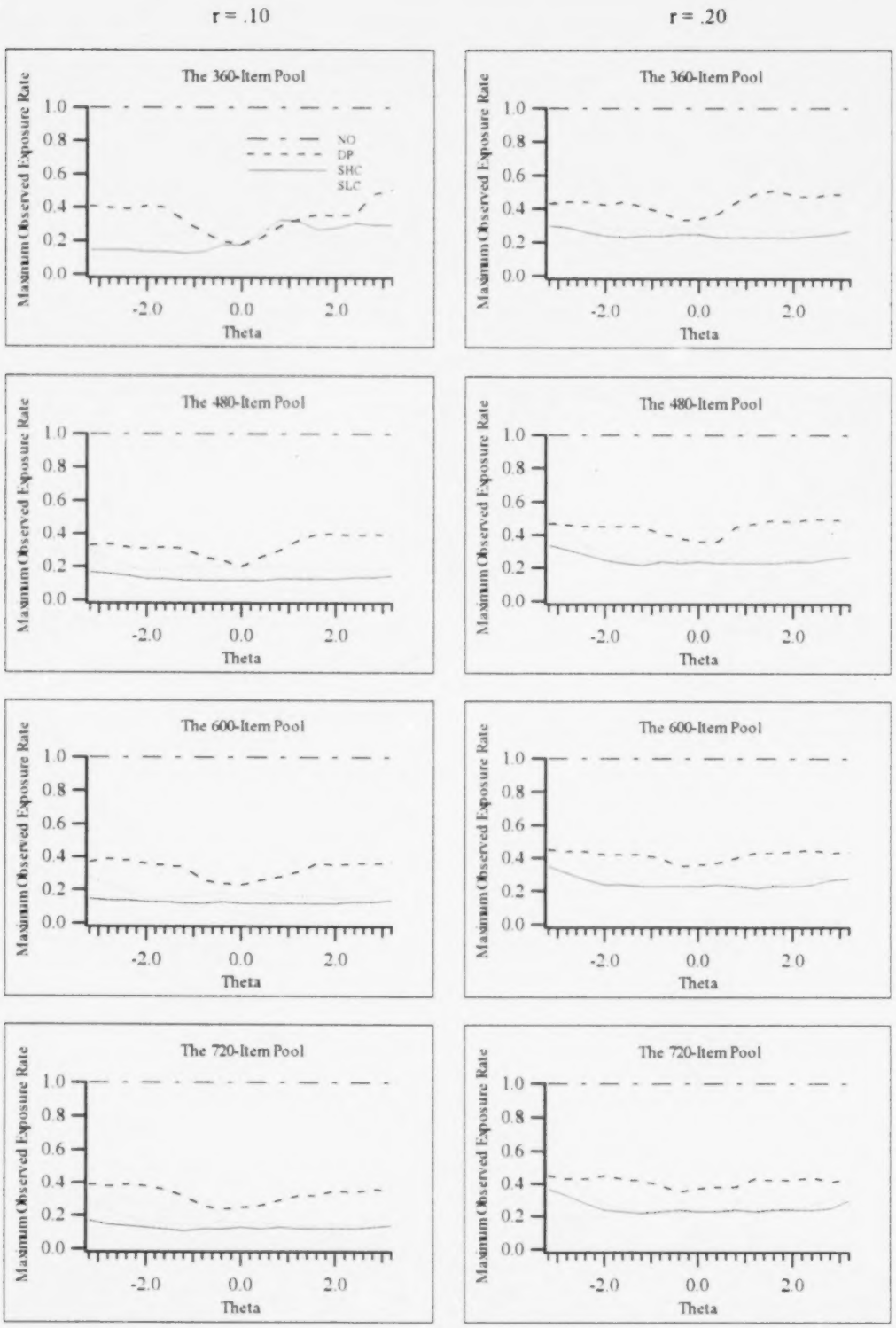

Figure 10. Conditional Maximum Observed Exposure Rates for the Various Conditions

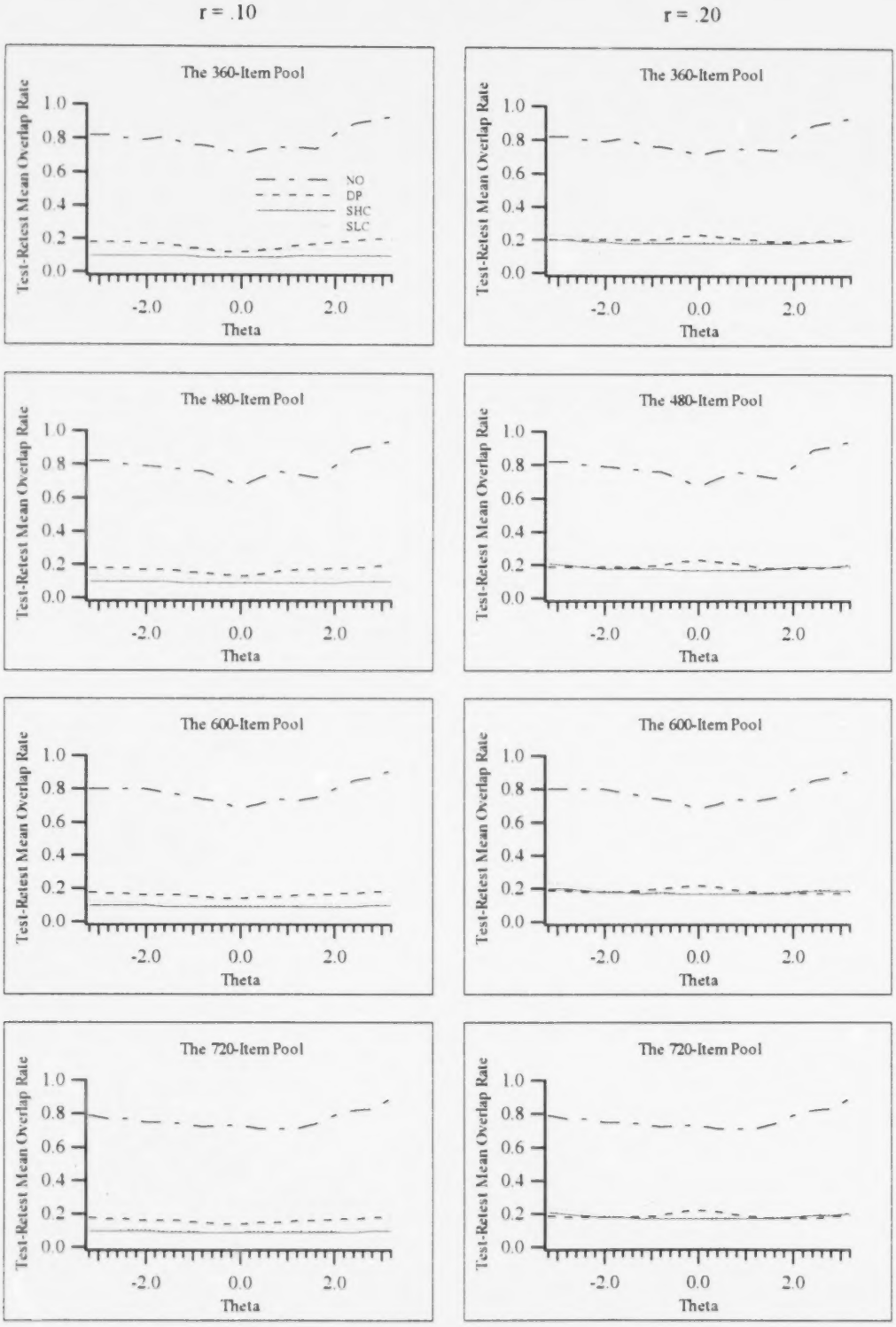

Figure 11. Test-Retest Mean Overlap Rates for the Various Conditions

BEST COPY AVAILABLE

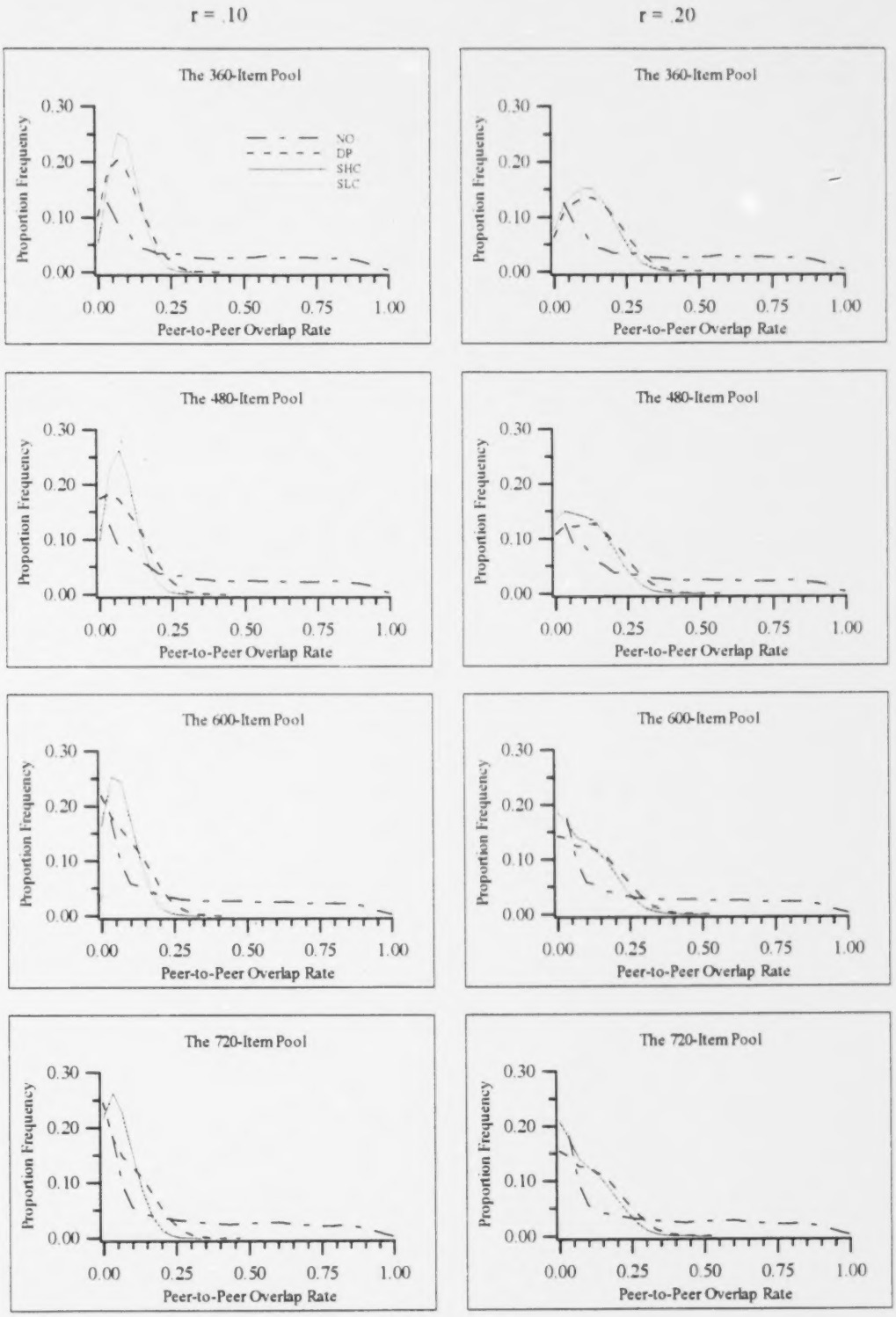

Figure 12. Full Distribution of the Peer-to-Peer Overlap Rates for the Various Conditions

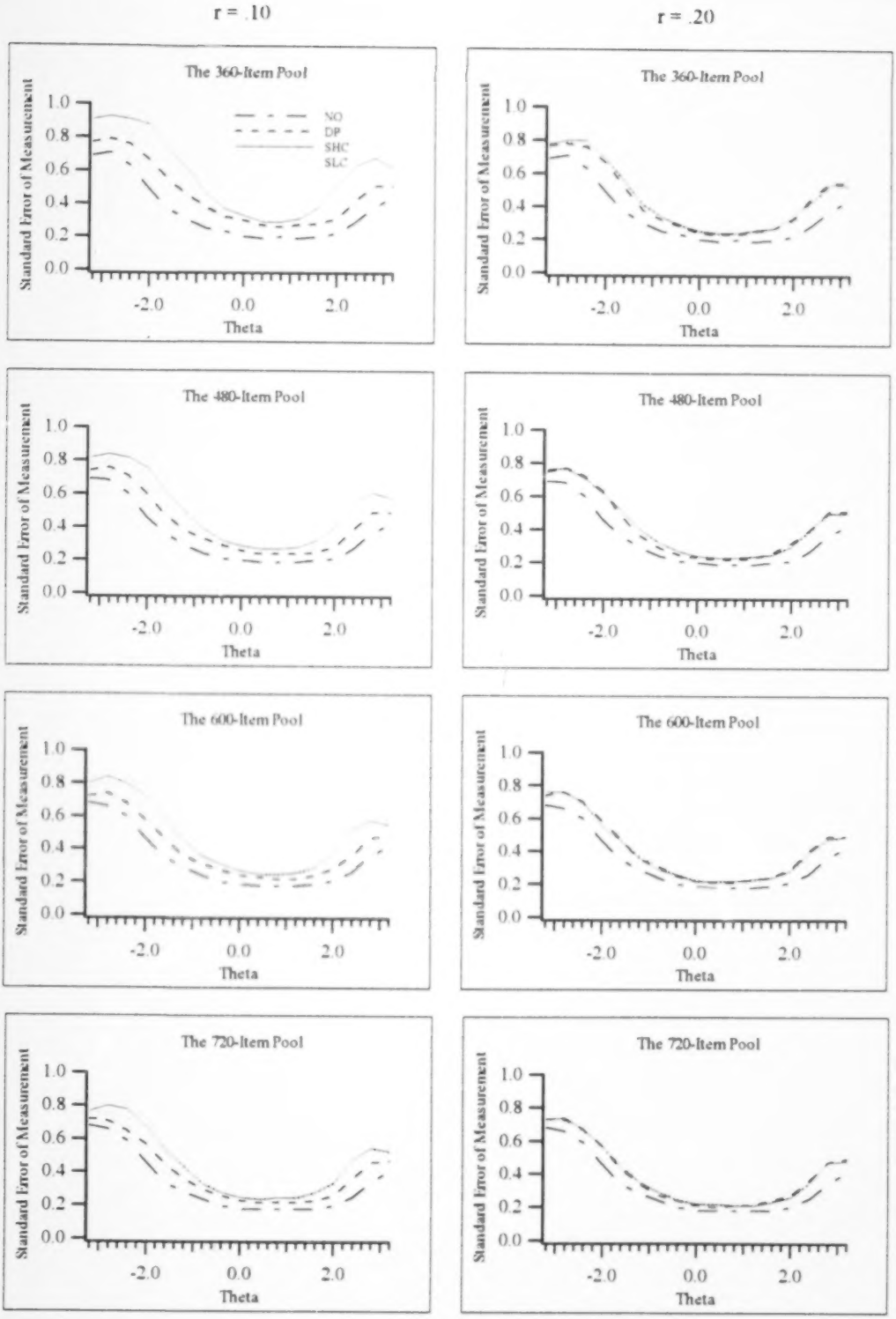

Figure 13. Standard Errors for the Various Conditions

BEST COPY AVAILABLE